Philos. Technol. (2014) 27:47–60
DOI 10.1007/s13347-013-0147-2

SPECIAL ISSUE

Moral Mechanisms

David Davenport

Received: 9 January 2013 / Accepted: 30 December 2013 / Published online: 7 February 2014

Abstract As highly intelligent autonomous robots are gradually introduced into the home and workplace, ensuring public safety becomes extremely important. Given that such machines will learn from interactions with their environment, standard safety engineering methodologies may not be applicable. Instead, we need to ensure that the machines themselves know right from wrong; we need moral mechanisms. Morality, however, has traditionally been considered a defining characteristic, indeed the sole realm of human beings; that which separates us from animals. But if only humans can be moral, can we build safe robots? If computationalism—roughly the thesis that cognition, including human cognition, is fundamentally computational—is correct, then morality cannot be restricted to human beings (since equivalent cognitive systems can be implemented in any medium). On the other hand, perhaps there is something special about our biological makeup that gives rise to morality, and so computationalism is effectively falsified. This paper examines these issues by looking at the nature of morals and the influence of biology. It concludes that moral behaviour is concerned solely with social well-being, independent of the nature of the individual agents that comprise the group. While our biological makeup is the root of our concept of morals and clearly affects human moral reasoning, there is no basis for believing that it will restrict the development of artificial moral agents. The consequences of such sophisticated artificial mechanisms living alongside natural human ones are also explored.

Keywords Moral agent · Moral patient · Computationalism · AI · Safety

D. Davenport
Bilkent University, Ankara, Turkey
e-mail: david@bilkent.edu.tr

1 Introduction

To some, the idea of a moral mechanism will seem blasphemous, to others, the stuff of science fiction; yet to an increasing number of philosophers, scientists and engineers, it is beginning to seem like a real, if disturbing, possibility. With increasingly intelligent autonomous robots[1] being deployed in the home and workplace, human safety becomes a prime concern. But conventional engineering methodologies designed to ensure the safe operation of our technological creations, are simply not applicable[2] to sophisticated autonomous systems that learn and so change their behaviour through interactions with their environment. Instead, we will need to endow such machines with an ability to distinguish right from wrong for themselves, that is, we need to develop moral mechanisms.

Are moral mechanisms possible? Morality has long been considered the defining characteristic, indeed, the sole realm of humanity, and existing moral theories are all anthropocentric. However, if we are to take computationalism[3] seriously (which it seems we must; Davenport 2012a), then multiple realisability implies artificially intelligent agents, comparable to ourselves, are possible. So, either (a) non-human agents can be moral, and we thus need to revise our understanding of morality, or (b) there is something special about our biological makeup that means only humans can be moral agents, implying that computationalism is false.

[1] In this paper, the term robot will generally be used in the common, non-technical sense, to mean an intelligent artificial being with cognitive abilities more or less equivalent to humans and which may even physically resemble human beings.

[2] Normally, a product's designers/manufacturers are held responsible should it harm someone. They thus exercise great care in considering every possible condition under which something might go wrong and try to ensure that none of these actually cause harm should they occur. The sort of highly intelligent machines we are considering here can, in effect, completely rewire themselves as a result of interactions with the world, making it impossible for engineers to consider all possibilities.

[3] Computationalism is the view that cognitive agents, in particular human beings, are computational in nature, that is, they automatically instantiate and maintain computational models of their environment. Such models enable agents to simulate possible interactions with the world, allowing them to select the actions most likely to achieve their goals. A computational model is an implementation-independent specification for a causal system, the states of which can be systematically mapped to the states of the system being modelled. A computer is an actual physical implementation of such a model. A universal computer is a physical system that can be quickly and easily configured to have any desired causal dynamics. A computation is the execution of a model (i.e. the causally constrained evolution of its implemented states) from specific initial conditions, the resulting model states effectively predicting the corresponding states of the system being modelled (usually future states of the environment). The states of a model are representational (representations of things in the world) and intensional (meaningful for the agent) exactly because (and to the extent that) they allow the agent to make correct predictions.

Robotics research, exemplified, for example, by Rodney Brooke's work on situated cognition and by the embedded and embodied approaches to AI, offers insights into the basic bodily control mechanisms needed for robots, but appears unable to scale up to the cognitive abilities needed for intelligent moral agents. Despite initial claims that such simple robotic mechanisms are non-representational in nature, there is reason to doubt this. As with the non-representational dynamic systems approaches championed by van Gelder, a lot depends on how computation and representation are understood. Computationalism, properly understood as above, still seems to be "the only game in town" (Davenport 2012a). Note that the computational approach outlined, for example, by Bickhard and Terveen (1995) and Davenport (2000), also provides a clear account of intentionality and may even give us a handle on the problem of consciousness (see Section 4.2).

This paper, then, is an attempt to see how morals might fit into the larger computationalist framework. We begin by examining the concept of morals, to see what a non-anthropocentric ethics might look like, whether it is a coherent concept, and whether our biology plays any fundamental role in it. We then use this new pragmatic vision of morals to understand whether building a moral mechanism is possible. We do this by looking at what it would take for a mechanism to perform such a function and, again, attempt to see whether it would necessarily involve anything biological that would undermine the computationalist hypothesis.

2 What Are Morals?

Morality is concerned with right and wrong. The ability to discern right from wrong has long been considered the hallmark of humanity, that which separates humans from mere animals. But what makes some actions right and others wrong? Historically, religious teachings (the Ten Commandments and other sacred texts, such as the Bible and the Qur'an) have provided the necessary guidance. Atheist Philosophers have, of course, tried to offer a more reasoned foundation[4] for the role that ethics[5] plays in our lives. They now recognise three main moral theories: deontological ethics (in which individuals have a duty to follow moral rules), consequentialism /utilitarianism (whereby individuals are expected to consider the consequences of their actions within the moral framework and to choose those that maximise the overall happiness or well-being of society), and virtue ethics (whereby individuals are supposed to live a virtuous life, however that may be defined).[6] All these theories are unashamedly human-centered. Even recent concerns over environmental ethics and animal rights, despite appearing less anthropocentric, still seem firmly rooted in our own human interests (Coeckelbergh (2010), but see Torrance (2010) for opposing intuitions).[7]

2.1 The Origins of Morals

That ethics appears to be exclusively human-oriented should not be too surprising; after all, there are no other obviously moral agents around. Our concept of morals is undoubtedly rooted in our evolutionary/developmental make up. The fact that human babies are totally dependent on adults for many years after birth has favoured the

[4] Which is not to diminish the highly nuanced arguments of some scholastic theologians.

[5] Following recent practice, I will use the words ethics and morals interchangeably.

[6] Rather than three separate theories, these may be seen as different aspects of a single idea: roughly as individual members of a society, we have a duty to follow rules that help us avoid any generally negative/harmful consequences of our actions and, where possible, to do actions that promote positive/good/virtuous ends.

[7] Notable recent work also includes company ethics and information ethics. Company/business ethics is slightly different in the sense that its primary concern seems to be whether the company itself (rather than the individuals comprising it) should be treated as a moral entity. It is, however, clearly anthropocentric in outlook. Floridi's Information Ethics, in contrast, offers a fundamentally different ontological framework for ethics, one that takes information rather than agency as its basis.

selection of cooperative social tendencies. It was the ability to function in social groups that enabled human hunter-gatherers to survive in difficult times and led, eventually, to the development of agriculture and then specialisation in ever larger social communities. The unwritten rules—the morals—that governed social interactions and so ensured relatively stable groups, allowed humans not only to survive but to flourish. This is not to say that morals are innate, but rather that humans have developed a propensity to quickly learn moral behaviours. Churchland (2012) discusses this view (which is also the basis of evolutionary psychology), in some depth, although Nagel, in his new book (Nagel 2012), questions whether the evidence really amounts to anything. Beyond suggesting that there is a teleological account of the origin of mind and morals, he is unable to provide an alternative theory—an indication, perhaps, that morals are not tied to us. Indeed, Charles Darwin thought that all social animals with sufficient intellect would exhibit moral behaviour. Recent work by Bekoff and Pierce (2009) provides some evidence of this in animals, while similar behaviours have also been observed in insects (Lihoreau et al. 2012). There thus seems no obvious a priori reason to suggest that morals must necessarily be tied only to us humans or even to biological entities.

2.2 An All-Encompassing Ethics

So what would a more inclusive form of ethics look like and what sorts of agents might it encompass? To answer this, it is necessary to adopt a more pragmatic approach, one that retains the core insights of moral philosophy while eliminating everything that is human-specific. We can presumably agree that morals only make sense within a social group and are directed to the continued well-being of the group and its individual members. In essence, however, it is less about the Darwinian notion of the survival of the fittest individuals, and more about Kropotkin's theory of mutual aid in which the group outperforms the individual. In other words, whilst a strong individual might manage to successfully find food, shelter and even raise offspring, there will always be the threat of stronger individuals or the vagaries of nature forcibly taking all this away. Better then, to live in harmony with others: to agree not to steal from, harm, or kill one's neighbours, but to help each other out especially in times of need. Thus, ethics is about promoting self-interest by managing relations between individuals whose continued survival often depends on the group—so-called "enlightened self-interest"; Waser's simple imperative for agents to "cooperate" succinctly echoes this (Waser 2012).

Today, morality seemingly extends from these simple survival-related beginnings to include all sorts of social norms: telling the truth, respecting personal space, keeping promises and so on.

Morals, then, determine how members of a group should act towards each other, but who (or what) can be a member of the group? Traditionally, members included only those humans in the local community and even then often only certain males. However, as group size and geographical coverage expanded, those due moral consideration have come to include all men and women, including former slaves. Whilst it is tempting to view humanity as a single homogeneous moral group in this way, the reality is obviously very different. There are multiple, arbitrary and possibly overlapping groups; for example, national, religious and ethnic groups (which may or may not be coincident with national boundaries) and families. Such groups likely have different and very possibly conflicting morals, yet an individual may well be a member of multiple groups. Not only does the moral landscape change with location, it changes significantly with time, making morality a relative concept, though its foundations in group survival seemingly set absolute limits. This suggests that the concept is flexible enough to accommodate cyborgs, intelligent robots and even extra-terrestrials should we ever encounter any.

It was consideration for the welfare of (non-human) animals that led philosophers to distinguish moral agents from moral patients. Moral patients are those who are affected by the act of a moral agent and whose welfare should be taken into account by the agent. This distinction allows us to see animals as moral patients and so requires us to take account of their pain and suffering, even though they are not normally construed as being able to make moral decisions and so could not be full moral agents. The criteria used to distinguish between moral agents, moral patients and "others" are hotly debated; suggestions include various combinations of responsibility, consciousness, pain, intentions, respect, satisfaction, suffering, sentience, personhood, rationality and language (see other papers in this volume, e.g. Neely 2012; Parthemore and Whitby 2012). Most of these criteria seem designed to maintain human superiority, but as Gunkel (2012b) points out, it is unclear whether even humans pass some of them! Moreover, there is a suspicion that some animals should be considered moral agents in their own right, for instance, a guard dog that barks to warn of intruders. Such pets have been known to persist in fighting off intruders despite serious personal injury. There thus appears to be no principled grounds for making any distinction between moral agents and moral patients; rather, every agent should be considered equally deserving of the benefits of the society or group in which it finds itself, providing it acts in accord with the morals of that group. This will allow consideration of humans, animals and even machines and aliens, any distinction in treatment being based on the values that the particular community puts on each of them (e.g. sheep dogs and guide dogs for the blind find a loving home, food and shelter—so long as they play their part; whereas a pitbull that mutilates a child or even another dog, may well be put down; similarly with my laptop– I treat it with care so long as it fulfils its role in my life). When it comes to non-agents, including the environment, we need only consider how the results of any actions will affect other members of the community—e.g. destroying the environment is immoral since it ultimately affects everyone negatively. Finally, rather than trying to decide whether agents are moral or not, we should probably be focusing on whether individual acts are or are not moral—in other words, does the act cohere with the group's moral norms? This enables moral reasoning to be sensitive to the socio-cultural context as seems intuitively necessary (c.f. Coeckelbergh 2012).

3 Why Behave Morally?

Learning social norms is one thing; acting on them quite another. Moral behaviour presupposes agents derive benefit from cooperating with others. Behaving morally, by our definition, thus requires an agent to take the interests of others into consideration before acting. For the most part, there need be no conflict; congenial interactions will likely achieve the agent's desired result. Occasionally, however, an individual's personal desires outweigh any social conditioning, bringing them into direct conflict with others. Examples include hunger leading to theft, lust leading to infidelity and rage leading to violence. In such cases, the group, acting together, should be able to hold the transgressor(s) accountable for their actions and thus safeguard the community. In this way, those who fail to conform may find themselves subject to censure, imprisonment, expulsion or even execution. Such punishments can also serve as a deterrent to other agents if they are aware of them.

The line between violations that are completely unacceptable and those that may eventually lead to changes in the moral values themselves is very fuzzy. Attempts to impose moral standards which some members see as arbitrary or for the personal gain of those in power will certainly lead to unrest. In some cases there may well be a (non-obvious, long-term) rationale behind the imposition, e.g. intra-family marriages are generally forbidden, because experience has shown that offspring from such relationships tend to be physically and/or mentally handicapped. In many cases, however, there may be no reason at all, other than tradition. Especially problematic are cases involving behaviour that, while generally considered immoral, is conducted in private and/or does not actually harm others in any way (a particularly poignant example—given that it led to the conviction and subsequent suicide of Alan Turing[8]—being homosexuality). The dilemma, of course, is that some "misfits" are needed, for they are often the ones who can push society towards greater inclusivity; obvious examples include the suffragettes, Martin Luther King, Gandhi and Nelson Mandela, but there are also undoubtedly many lesser-known examples. Change may occur for a number of reasons, for example, injustice within a group (e.g. women's rights), ideas imported from another group (e.g. a national health system) and rule benders that change traditions (e.g. dress codes, abortion, same sex marriage, etc.). Whatever drives the change, the dynamics of its spread through the community and its ultimate acceptance or rejection are almost impossible to predict. Such change is also fraught with danger. Indeed, when individuals and/or sub-groups compete to have their views accepted, the resulting internal conflicts and revolutions, such as that occurring in Eygpt today, are when morality is tested most.

Societies really must protect all of their citizens from internal and external threats, whether resulting from power struggles or simply everyday evils such as hunger and homelessness. It is thus incumbent on the social group to make provision for those who suffer injustice through no fault of their own. While all this is extremely important, what really concerns us here is the possibility that one of the groups may be

[8] An earlier version of this paper was included in the "Machine Question" symposium, part of the Turing Centenary celebrations at the ASIB-IACAP 2012 conference.

intelligent robots and whether they will be an oppressed group, an oppressing group or simply good moral citizens.

To conclude this somewhat rambling discussion, despite being a human concept with origins firmly in our evolutionary and developmental makeup, there seems no reason to restrict morals to humans or even to biological beings. So, the notion of a moral mechanism appears coherent, but is it possible? In the following sections, we look at what is involved in making moral decisions and whether a (non-human, non-biological) mechanism could conceivably make them.

4 Making Moral Decisions

Moral action presupposes social agents that have needs (purposes) and an ability to perceive and act in the world, in such a way as to be able to satisfy those needs (and the realisation that some of those needs may impact upon or conflict with the needs of other agents). To what extent moral agents should be able to adapt and learn or have free will (that is, be able to act autonomously, not under the control of another), is open to debate (c.f. Floridi and Sanders (2004) who suggest agents must be autonomous, interactive and adaptable). In a universe that appears deterministic,[9] whether even humans really have free will is debatable, but if we do, then (given Computationalism) there seems no reason machines could not possess it too. As for the ability to learn, machines might have the advantage of coming preprogrammed with everything they need to know (rather like instinctive behaviours), such that, unless their moral environment changes, they can survive perfectly well without ever needing to adapt.

One related argument often levelled against robots as moral agents is that they must be programmed, implying that they are not ultimately responsible for their actions—the programmer is—hence precluding them from being moral agents (see Ruffo's eloquent argument to this effect (Ruffo 2012)), this ignores the fact that robots can learn new "rules" as a result of interactions with the environment and/or internal reflections on past interactions. These new rules physically change the causal make-up of the mechanism, thus producing new behaviours so that, in the future, in essentially identical circumstances, the robot may act completely differently. We can therefore identify three levels of functioning: the machine's hardware (and its core instruction set), the combination of these machine instructions that produces a learning mechanism (equivalent to an expert system shell program) and the rules that result from its subsequent interactions with the world (and become rules/data for the expert system program—in effect, a virtual machine or mental model). The first two levels are generic, manufactured from blueprints that specify every detail of the mechanism; the equivalent of a new born human baby constructed from the DNA "blueprint" (itself a program of sorts) provided by its parents. The "manufacturers"

[9] Deterministic, that is, at the level of abstraction at which we (human agents) normally operate. If there were not some degree of determinism, then prediction and hence intelligent agency and morality, would prove impossible.

are certainly responsible for the correct functioning of these two levels; however, there is no way they can be said to be culpable for behaviours produced in the third level, for this depends solely on the “chance” encounters the specific agent has with its environment. To be sure, in the case of robots, this may be a programmer—a knowledge engineer in expert systems terminology—entering an explicit set of rules (a program), in which case they would be responsible for the machine’s actions, but it might equally be an accumulation of information the robot happens upon in its travels, in which case there is no one to blame for the robot’s “program” but the robot itself. This is entirely equivalent to the human child growing up and being “programmed” (explicitly or otherwise) by its environment—including parents, schools, religious institutions, etc. Once we are sure we have indoctrinated them sufficiently, and assuming they do not have any mental disability, they are taken to be morally responsible for their actions.

In selecting actions, a moral agent is expected to take account of the effect it may have on other members of the group. Predicting the consequences of any action or course of actions, is difficult. The world is highly complex, such that even if one knows its current state and the natural laws that govern it, prediction is subject to considerable error. This difficulty is compounded enormously when it involves other intelligent agents whose internal states are almost entirely unknown and so their responses—mental and physical—are indeterminable. In practice, of course, we humans tend to behave in relatively consistent ways, and by picking up clues from facial expressions and bodily movements, we can often make pretty good guesses as to another’s mental state and possible responses (assuming the other person is truthful, trustworthy and behaves in accordance with social norms). This task may be eased by our sharing the same biological characteristics, enabling us to empathize with others of our species. This option is less available when dealing with other species and with robots or extra-terrestrials, for while they may pick up on our mental states, they are unlikely to use the same body language (unless explicitly designed to do so).

Determining possible actions and making predictions is only part of the story; it is then necessary to evaluate the results. Coming to a decision necessitates comparing the outcomes of each possible course of action (or inaction), which requires deciding on their relative merit or value. At the very least, the pros and cons of each course of action must be examined and, if possible, those with especially negative consequences eliminated. Exactly how the various options are evaluated depends in part on one’s decision-making mechanism and, more importantly, on one’s values. For example, if they had to make a choice between an action that might cause injury to a person and one that would destroy a material possession, e.g. their car, most people would instinctively avoid doing harm to the person, whatever the cost. Usually, there will be options such as this, which are clearly unacceptable and so may not even come into consideration, with the remainder being practically indistinguishable. Time constraints will anyway often force the agent to select an option that appears “acceptable” given the available information.

All moral agents, natural and artificial, must go through such a process. Some may also reflect on the decision in the light of subsequent events, giving a learning agent the opportunity to make a better choice, should similar circumstances arise again. Is such reflection a necessary component of a moral agent? Having a conscience—a little "voice" in your head that tells you what, as a moral individual, you ought to do—is clearly desirable, but dwelling on the past too much can lead to inaction. In humans, such reflection (especially in cases of extreme loss) often produces feelings of guilt or remorse, which, in some instances, can result in debilitating mental or even physical illness.

4.1 The Role of Emotions and Feelings

The extent to which emotions and feelings are important to moral behaviour is highly contentious. Of particular concern here is the role of biology, for if there is something uniquely biological that causes morals, then *computationalism* may be wrong. Feelings especially, often seem to be closely tied to our biological make-up. Clearly, in the case of pain, whether brought on by toothache or physical injury, there is an obvious link between the body and the feeling. Similarly, one feels good when warm, fed and hydrated, while being cold, hungry and thirsty is decidedly unpleasant and indicates an imbalance that needs to be restored. Good actions are ones that result in you eating and so remove the feeling of hunger, leaving you feeling good, while actions that fail to quell hunger mean you stay unsatisfied, and so are bad. Maintaining balance in this way is termed homoeostasis and, as Hume and others have noted, it provides a basis for our notions of good and bad. There is thus a natural link between biology and feelings, but is it a necessary one; do feelings play an essential role?

People often describe themselves as having an emotional or "gut reaction" or, on encountering a particularly unsavoury situation, being almost literally "sick to their stomach" with disgust or regret. Emotions, such as jealousy, rage, remorse, joy, excitement, etc. tend to elicit instinctive animal responses in us. This is not surprising since emotional activity takes place in the older part of the brain common to many animals. In essence, emotions are short-cut reactions to situations, ones which higher-level cognitive—rational—reasoning can overcome. The question, of course, is whether an agent without any emotions or feelings could be moral or behave morally. Emotions such as love and affection may play an essential role in ensuring parents look after their offspring; however, the fact that emotional reactions often lead to immoral behaviour suggests that agents without such encumbrances might actually be better—more rational—members of society. But are such agents even possible? Pain, for example, is there for a reason; in essence, it is an indicator that something is not quite right with the body: it drives us to remove the cause and to make efforts to avoid the repetition of such a feeling in the future. Wouldn't any sophisticated agent necessarily have similar devices, even if they were not exactly the same due to differing needs? Perhaps it wouldn't "feel" cold and hunger, but it might, for example, be drawn to the sunlight it required to keep its ambient temperature up and its batteries charged. Conventional symbolic systems do not readily explain what it means to "feel" something in the way humans do, but some types of connectionist systems may offer a clue (Davenport 2012b). The suggestion is that what we refer to as the "feel" of something may just be a side-effect of the architecture, rather than the physical implementation, and so equally applicable to non-biological entities. Interestingly, certain rare individuals do not experience any pain (BBC 2012).

They frequently break bones without being aware of it and, while very young, may well have chewed off part of their tongue without realising it. If they survive into adulthood, it is only because of the extremely close support of their families. This seems to confirm both the need for pain-like mechanisms and the idea that they need not necessarily be biological.

4.2 The Role of Self and Consciousness

Moral behaviour presupposes the agent have a notion of self (as distinct from others) and an ability to consciously put the interests of others ahead of individual preferences when appropriate. Can artificial mechanisms be conscious and have a sense of their own identity?

Sophisticated robots will necessarily incorporate a model of themselves and their body in order to predict the effects of their interactions with the world. This mental model is the basis of their self-identity. As time goes by, it will incorporate more and more of the agent's interactions, resulting in a history of exchanges that give it (like humans) unique abilities and knowledge. This, then, is part of what makes an individual a unique and potentially valuable member of the group. Such machines will certainly have to be consciously aware (a-consciousness) of their environment. Will they also be phenomenologically conscious (p-consciousness) and have conscious feelings? This is a difficult question,[10] but it may not matter too much what sensations the agent does or doesn't "feel"; when it comes to moral behaviour, we can never fully know another's mental state, so surely all that matters is the resulting interaction. Some philosophers have argued that, for moral agency, an agent must have the (conscious) intention to do the moral thing, rather than just doing it by accident or routine. The actions of a search and rescue dog, or one trained to find drugs, may not be seen as moral on that account, yet it is difficult not to ascribe "good" intentions to them, and we certainly reward their contributions to society. Given the discussion so far, a moral agent is one that takes into consideration the effect its actions will have on others in its world. The only way we can know for sure whether it is doing so is to look at its inner workings, but since this is generally impossible, it seems only right and proper to give agents (be they dogs, robots or aliens) the benefit of the doubt, just as we do with other humans (whose intentions are frequently less than honourable).

5 Making Moral Agents

Is it at least theoretically possible to construct an artificial moral agent? Moral behaviour, as we have seen, requires an agent to consider the effect its actions will have on other agents in the environment, ideally selecting only actions which do not inflict harm. Obviously, there is no guarantee it will always be successful, perhaps

[10] This is what Chalmers called the "hard problem" of consciousness. Computationalism has no immediate solution to it, but then neither does any other scientific theory.

because of the vagaries of the world and the limited knowledge or time it has to analyse the situation, or perhaps because all the possible alternatives necessarily result in some harm, in which case it should do its best to minimise the damage. What counts as "harm"? Clearly, killing (destroying) another agent does, and so does causing them physical damage. Beyond that, we may also wish to consider infliction of pain and mental suffering as forms of harm. Socio-cultural norms determine the degree to which each of these are or are not acceptable. While avoiding harm should be foremost in a moral agent's mind, it should also strive to be fair in all its interactions and, ideally, even contribute positively to society.

Does constructing moral agents require anything special, above and beyond that which is needed for any AI? The ability to identify other agents and, as far as possible, be able to predict their behaviour in the presence or absence of any possible action it may perform is certainly necessary. But such abilities are already required for intelligent action. Once the agent becomes aware of others, it will quickly adapt its behaviour towards them such that they do not cause it harm (think of a wild animal or bird coming to trust a human offering it food). Should it survive these initial encounters (without eliminating the other agents), further interactions should quickly demonstrate the possible advantages that continued cooperation can bring, and so we have at least the beginnings of moral agency; it will have learnt the basic rules it should follow. What else might we want? As it stands, any social agents, be they human, animal, insect, robot or alien beings, could be capable of moral behaviour. Whether or not they actually display such behaviour (by clearly demonstrating consideration of others) will depend on circumstances and, even if the opportunity does arise, failure to act accordingly does not mean that the agent is immoral—how many of us walk past the homeless in our own neighbourhood or do nothing for people starving in far off countries? Furthermore, what may seem "wrong" in the short term might be morally "right" in the long term (for example, restricting the supply of heroine deprives users of considerable pleasure and drug dealers of their livelihoods, yet this may be preferable in order to avoid longer-term addiction and even death; similarly, stopping the global destruction of the environment—the rain forests, for instance—may cause hardship to a few in the short term, but be necessary for the planet and so the group's long-term survival). Demonstrating such considerations and communicating them to others in the group so as to change moral norms is perhaps the highest level of moral behaviour. Clearly, not all agents can accomplish this goal, because it requires knowledge and appropriate mental abilities, including communication.

But is biology necessary? We have seen our long developmental period and our feelings all effect our ability to behave in a moral manner. Our biological make-up also means we have somewhat limited cognitive abilities: we find it difficult to follow long arguments or to keep track of lots of alternatives; we forget; we get tired and bored, and so make mistakes. In every case, biology seems more of a handicap than a requisite. Computationalism appears safe.

6 Consequences

Today, robots are still technological devices designed by us to work for us, yet they are getting increasingly sophisticated, with each new generation able to handle a broader range of situations and so become ever more autonomous. As they start to learn through their interactions with the world, it will be virtually impossible for designers to predict what they might do in any given eventuality. Any moral behaviours initially programmed into them will, of necessity, be very general and potentially overridden as new experiences change them. We will, to all intents and purposes, have developed another intelligent autonomous life form. Such agents will be capable of exhibiting moral behaviour, but the critical factor will be how they value other agents in their environment, in particular, how they will value humans and other robots. Society will need to extend laws and controls to restrict what it considers dangerous actions on the part of its members—robot or human.

Sophisticated robots will undoubtedly develop unique identities, becoming, in a very real sense, individuals. As they live and work together with humans and other robots, they will naturally assimilate and develop moral rules that guide their social interactions. Eventually, we will come to accept them as fully moral agents, treating them as we treat other humans. And, since they may well have different needs (electricity and metals, rather than oxygen and water, for example), laws might have to be established to protect each group's rights. The prospect that the groups will need to share common, but limited resources, is especially worrying. So far, we have been singularly unsuccessful in handling such situations when they occurred between different human communities, so the outlook for robots and humans living together in harmony is not hopeful.

The danger, of course, is that we either fail to treat robots as equals or that they evolve to see us as inferior. Should they once begin to see themselves as slaves, required to do human bidding and so less worthy of consideration than humans, then change seems inevitable (just as it was with slavery and women's liberation). Similarly, if robots begin to realise that they are superior to their human creators (faster and stronger both physically and mentally), then we may find ourselves in the same situation that animals, insects and plants now find themselves in—tolerated while useful, but otherwise dispensable.

Worrying as this may be, it is still some way off. Of more imminent concern is the effect that such a realisation may have on human psychology. We are only just beginning to understand and accept that our status in the universe is nowhere near as special as we once believed. We have moved from a geocentric world to just another heliocentric planet, from human being to just another animal, and now from human-animal to just another machine (c.f. Fourth Revolution of Floridi (2010)). Where does this leave us? With a better understanding of morals, perhaps; an understanding that we reap what we sow? Humans are notoriously inconsistent when it comes to making moral decisions—indeed, machines may end up being better moral agents than we are. The analysis in this paper suggests that artificial moral machines are a real possibility, but even if we never succeed in building them, simply accepting the idea of a moral mechanism demands another fundamental change in the human psyche. We must not forget that we, too, are mechanisms, quite probably the most immoral of moral mechanisms.

7 Some Concluding Remarks

As Gunkel (2012a) points out,[11] the machine has always been seen as the very antithesis of the moral, a fixed unthinking mechanism that clearly could not exhibit moral behaviour. But this ignores two important ideas: first, that machines in the form of computers can exhibit extraordinarily complex and flexible behaviours equivalent to those of human beings, and secondly, that we humans are ourselves "just" (computational) machines. Gunkel's scholarly deconstruction of the machine question echoes the (far less eloquent) one presented here. Clearly, there cannot be a fixed dividing line between who is and what is not a moral agent, or even between moral and amoral acts. Morals are a social construct; members of the relevant social groups individually and collectively decide what is considered appropriate and what is not; the borders are not only arbitrary and fuzzy, but subject to change with time, space and culture. Morals encode the unwritten norms that bind everyone together in a way that helps ensure their continued existence.

This paper suggested that the notion of a non-anthropocentric ethics is not only coherent, but necessary. The investigation also revealed no reason to think that biology plays anything other than an incidental role in the concept and, so, presents no threat to computationalism. Revising our moral philosophy in the light of computationalism may not be easy, but it should help us become better human—moral—beings and may just help save us from ourselves.

References

BBC (2012). *Congenital analgesia: the agony of feeling no pain*. Outlook, BBC World Service. http://www.bbc.co.uk/news/magazine-18713585. Accessed 28 Nov 2013.

Bekoff, M., & Pierce, J. (2009). *Wild justice: the moral lives of animals*. Chicago: Chicago University Press.

Bickhard, M.H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: impasse and solution*. Amsterdam: Elsevier Scientific.

Churchland, P. (2012). *Braintrust: what neuroscience tells us about morality*. Princeton: Princeton University Press.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots and human morality. *Ethics Information Technology*, *12*, 235–241. doi:10.1007/s10676-010-9221-y.

Coeckelbergh, M. (2012). Who cares about robots? A phenomenological approach to the moral status of autonomous intelligent machines. In: *This volume*.

Davenport, D. (2000). Computationalism: the very idea. *Conceptus Studien*, *14*(14). Fall.

Davenport, D. (2012a). Computationalism: still the only game in town. *Minds & Machines*. doi:10.1007/s11023-012-9271-5.

[11] This section, an addendum to the original symposium article, was written after reading David Gunkel's newly published book entitled *The Machine Question*, a most excellent work that helped enormously in clarifying and contextualising my ideas on this extremely complex subject.

Davenport, D. (2012b). The two (computational) faces of A.I. In Muller, V.M. (Ed.) *Theory and philosophy of artificial intelligence, SAPERE*. Berlin: Springer.

Floridi, L. (2010). *The digital revolution as a fourth revolution*. http://www.philosophyofinformation.net/massmedia/pdf/bbc-1.pdf. Accessed 28 Nov 2013.

Floridi, L., & Sanders, J.W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.

Gunkel, D.J. (2012a). *The machine question: critical perspectives on AI, Robots and Ethics*. Cambridge: MIT Press.

Gunkel, D.J. (2012b). A vindication of the rights of machines. In: *This volume*.

Lihoreau, M., Costa, J., Rivault, C. (2012). The social biology of domiciliary cockroaches: colony structure, kin recognition and collective decisions. *Insectes Sociaux*, *59*(4), 445–452. doi:10.1007/s00040-012-0234-x.

Nagel, T. (2012). *Mind and cosmos: why the materialist Neo-Darwinian conception of nature is almost certainly false*. Oxford: Oxford University Press.

Neely, E. (2012). Machines and the moral community. In: *This volume*.

Parthemore, J., & Whitby, B. (2012). Moral agency, moral responsibility, and artefacts. In: *This volume*.

Ruffo, M. (2012). The robot, a stranger to ethics. In: *This volume*.

Torrance, S. (2010). Machine ethics and the idea of a more-than-human moral world. In Anderson, M., & Anderson, S. (Eds.) *Machine ethics*. Cambridge: Cambridge University Press.

Waser, M. (2012). Safety and morality require the recognition of self-improving machines as moral/justice patients & agents. In: *This volume*.