Moral Mechanisms

David Davenport[1]

Abstract. Moral philosophies are arguably all anthropocentric and so fundamentally concerned with biological mechanisms. Computationalism, on the other hand, sees biology as just one possible implementation medium. Can non-human, non-biological agents be moral? This paper looks at the nature of morals, at what is necessary for a mechanism to make moral decisions, and at the impact biology might have on the process. 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 biology certainly affects human moral reasoning, it in no way restricts the development of artificial moral agents. The consequences of sophisticated artifical mechanisms living with natural human ones is also explored. While the prospects for peaceful coexistence are not particularly good, it is the realisation that humans no longer occupy a privileged place in the world, that is likely to be the most disconcerting. Computationalism implies we are mechanisms; probably the most immoral of moral mechanisms.

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. Existing moral theories are arguably all anthropocentric. However, if we take computationalism seriously (which it seems we must [3]), then multiple realisability implies artificially intelligent agents, comparable to ourselves, are possible. Can such non-human agents be moral or is there something fundamentally human about morality? To what extent, if any, does biology impact moral behaviour? Indeed, what exactly is moral behaviour, what would it take for a mechanism to exhibit it, and why does it matter? This paper examines these questions and outlines some of the consequences: philosophical, psychological and social. It is my attempt to make sense of the vast literature on the subject, and see how morals might fit into the larger computationalist framework. Given the increasing pace of research and development into robotics, a clear understanding seems essential. We begin, then, by developing a pragmatic understanding of the function of morality, then focus on the possibility of moral mechanisms and on the extent to which biology is relevant.

2 WHAT ARE MORALS?

Morality is concerned with right and wrong. The ability to discern right from wrong is often considered the hallmark of humanity; that which separates humans from mere animals. But what makes some actions right and others wrong? Historically, religious teachings (from the Ten Commandments[2] and sacred texts, such as the Bible and the Qur'an) have provided the necessary guidance. Philosophers, of course, have tried to offer a more reasoned understanding of the role that ethics[3] 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). All these theories are unashamedly human-centered. Even recent concerns with animal rights and environmental ethics, despite appearing less anthropocentric, are still firmly rooted in our interest in the survival of the human population ([2], but see [7] for opposing intuitions).

That work on ethics appears to be exclusively human-oriented should not be too surprising; after all, there are no other obviously moral agents around. Charles Darwin suggested that all social animals with sufficient intellect would exhibit moral behaviour.Recent work by Bekoff and Pierce [1] provides evidence of moral behaviour in animals, albeit somewhat limited, while similar behaviours have also been observed in insects [6]. It seems that artificially intelligent robots with intellectual capacities approximating our own may soon be a reality. The fact that such entities may be deployed, not only on the battlefield, but in everyday situations around the home and workplace, where they must interact with humans, make it essential that we understand the ethical issues involved.

So what would a more inclusive form of ethics look like and what sorts of mechanisms 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 children, there will always be the threat of stronger individuals 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. Ethics, then, is about promoting self-interest by managing relations between individuals whose continued survival depends on the group—so-called

[1] Bilkent University, Ankara 06800 - TURKEY, email: david@bilkent.edu.tr

[2] According to the wikipedia entry, the Ten Commandments may have been based on much older laws from the Hitit empire that occupied Central Anatolia–Ankara–and extended as far as Egypt, circa 2000BC.

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

"enlightened self-interest".

Morals, today, seemingly extend from these simple beginnings to include all sorts of social norms: telling the truth, respecting personal space, limited touching, keeping promises, and so on. Of course, such rules and conventions must be learnt. Human children usually learn from their parents and by playing with other children in the relatively safe confines of the home, as well as from religious teachings and school.

3 WHY BEHAVE MORALLY?

Learning social norms is one thing, acting on them quite another. Behaving morally, almost by definition, requires an agent to put the interests of others ahead of its own individual preferences (or at the very least 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 desired result. In other words, we can usually get what we want by playing the social/moral game. 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, will always be able to overcome/restrain the "rogue" individual. In this way, those that fail to conform may find themselves subject to censure, imprisonment, or even death. Much philosophical discussion has centered around the "social contract" that individuals seem to implicitly sign up to when they are born into a community, and whether society has the right to enforce compliance, given that the individual did not make a conscious choice to join and is usually unable to leave. There is certainly a danger if society attempts to impose moral standards which its members see as arbitrary or for the personal gain of those in power. 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 done 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—being homosexuality). The dilemma, of course, is that society really does need some "rogues", for they are often the ones who can change it for the better; obvious examples include the suffragettes, Martin Luther King, Gandhi, and Nelson Mandela. At the same time, society has a duty to protect its individual citizens, not only from external threats, but from everyday evils such as hunger. For this reason, some sort of supportive, welfare state is needed. Society must make provision for those who suffer injustice through no fault of their own, whether the result of financial difficulties brought about by failures of Capitalism, or because of failures in the law, leading to innocent persons being wrongfully imprisoned. While all this is extremely important, in what follows, we will be more concerned with the moral decision making process and what effect biology may have on it.

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 their needs. To what extent they should be able to adapt/learn, or have free will (that is, be able to act autonomously, not be under the control of another), is open to debate (c.f. Floridi & Sanders, who suggest agents must be autonomous, interactive and adaptable). In a universe that looks deterministic, 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 behaviour), such that, unless their (cultural/normative) environment changes, they can survive perfectly well without ever needing to adapt.

In selecting its actions, the moral agent is expected to take account of the effect this 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, prediction may be subject to considerable error. This difficulty is compounded enormously when it involves other intelligent agents whose internal (mental) states may be completely unknown and so their responses 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 empathise with others (perhaps aided by so-called mirror neurons). This option is less available when dealing with other species and robots, for while they may pick up on our mental states, they are unlikely to send out signals in a similar way (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 moral theory and, more importantly, on one's values. For example, if they had to make a choice between a 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, whilst the remainder being practically indistinguishable. Time constraints will anyway often force the agent to select an option that appears "acceptable" given the available information. Of course, subsequent events may show it was far from the optimal choice, but by then it is too late.

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 in the future, 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 a good thing, but dwelling on the past too much can be counterproductive. In humans, such reflection (especially in cases of extreme loss) often produces feelings of guilt or remorse, which, in some instances, can result in mental or even physical illness.

4.1 The role of emotions & 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. 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 satisfy your hunger, mean you stay unhappy, and so are bad/undesirable. Maintaining balance in this way is termed homoeostasis. There is thus a natural link between biology and feelings, but is it a necessary one?

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. 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 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 and so drives us to remove the cause and to make efforts to avoid 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" hunger, but it might, for example, be "uncomfortable" out of the sunlight it required to keep its batteries charged. Conventional symbolic systems do not readily explain what it means to "feel" something, but some sorts of connectionist systems may offer a clue [4]. 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.

4.2 The role of self & consciousness

Moral behaviour presupposes a notion of self 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 model themselves in order to predict the effect their actions will have on the world. This model is the basis of their self identity. As time goes by, it will incorporate more and more of the agent's interactions with the world, resulting in a history of exchanges that give it (like humans) unique abilities and knowledge. This, then, is part of what makes an individual, an individual and a 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, 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 really know another's mental state, so surely all that matters is the resulting interaction. Some philsophers 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 to not to ascribe "good" intentions to them, and we certainly reward their contributions to society.

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 behaviour 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 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. One might add that it should try to be fair in all its interactions and to contribute positively to society, but such characteristics may be too much to expect.

Does constructing moral agents require anything special, above and beyond that needed for any AI? The ability to identify other agents and, as far as possible, be able to predict their behaviour in the presence and absence of any possible action it may perform, is certainly needed. 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/norms it should follow. What else might we want? As it stands, any social agents—be they human, animal, insect, robot or alien beings—would seem capable of moral behaviour. Whether or not they actually display such behaviour (by clearly putting social needs ahead of their own), will depend on circumstances and, even if the opportunity does arise, failure to act accordingly does not mean the agent is not generally moral—how many of us walk past the homeless in our own neighbourhood or do nothing for those starving in far off countries?

Is biology necessary? The fact that human babies are so weak and helpless when they are first born, means they cannot harm others. Their total reliance on their parents naturally encourages the development of cooperative tendencies, which, again, are the first steps towards moral behaviour. As they grow, they become stronger and more independent, and increasingly test the limits of their parents, siblings, teachers and friends. Hopefully, they emerge from this formational period with a reasonable understanding of right and wrong (and the huge grey area between). It is only after children have developed sufficiently (mentally, as well as physically), that they become legally responsible for their actions (for example, in many countries juveniles cannot be sent to prison, even for murder). Given that robots may be physically very strong and so dangerous from the moment they "come alive", we may need some way to ensure they are also "born" with the relevant moral experiences. What experiences are needed and how they can be encoded and enforced is obviously an important question, not just for artificial moral mechanisms, but for human ones too.

Our long developmental period and our feelings and emotions, 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. Here again, then, biology seems more of a handicap than something essential.

6 CONSEQUENCES

Today, robots are still technological devices, designed by us to work for us, yet they are getting increasingly sophisticated, each new generation being able to handle a broader range of situations and so becoming ever more autonomous. As they start to learn through their interactions with the world, it will be virtually impossible for designers to be able to predict what they might do in any given situation. Any moral behaviours initially programmed into them will, of necessity, be very general and potentially overridden as new experiences change it. We will, to all intents and purposes, have developed another intelligent autonomous life form. Such agents will be capable of exhibiting moral behaviour, the deciding factor is how they value other agents in their environment; in particular, how they will value humans and other robots. Society will 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 incorporate/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 at all good.

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 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 now find themselves in—tolerated while useful, but otherwise dispensable.

Worrying as this may be, it is still a long 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. Floridi's Fourth Revolution [5]). 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; probably the most immoral of moral mechanisms.

REFERENCES

[1] M. Bekoff and J. Pierce, *Wild Justice: The Moral Lives of Animals*, Chicago University Press, 2009.

[2] M. Coeckelbergh, 'Moral appearances: emotions, robots and human morality', *Ethics Information Technology*, **12**, 235–241, (2010).

[3] D. Davenport, 'Computationalism: Still the only game in town', *Minds & Machines*, (2012).

[4] D. Davenport, 'The two (computational) faces of ai', in *Theory and Philosophy of Artifcial Intelligence*, ed., V. M. Muller, SAPERE, Berlin: Springer., (2012).

[5] L. Floridi. The digital revolution as a fourth revolution, 2010.

[6] M. Lihoreau, J. Costa, and C. Rivault, 'The social biology of domiciliary cockroaches: colony structure, kin recognition and collective decisions', *Insectes Sociaux*, 1–8, (2012). 10.1007/s00040-012-0234-x.

[7] S. Torrance, 'Machine ethics and the idea of a more-than-human moral world', in *Machine Ethics*, eds., M. Anderson and S. Anderson, Cambridge University Press, (2010).