J. Expt. Theor. Artif. Intell. 12(2000)251–261

Prolegomena to any future artificial moral agent

COLIN ALLEN, GARY VARNER and JASON ZINSER

Department of Philosophy, Texas A&M University, College Station, TX 77843-4237, USA
email: colin-allen@tamu.edu; g-varner@tamu.edu; j-zinser@philosophy.tamu.edu

Abstract. As artificial intelligence moves ever closer to the goal of producing fully autonomous agents, the question of how to design and implement an artificial moral agent (AMA) becomes increasingly pressing. Robots possessing autonomous capacities to do things that are useful to humans will also have the capacity to do things that are harmful to humans and other sentient beings. Theoretical challenges to developing artificial moral agents result both from controversies among ethicists about moral theory itself, and from computational limits to the implementation of such theories. In this paper the ethical disputes are surveyed, the possibility of a 'moral Turing Test' is considered and the computational difficulties accompanying the different types of approach are assessed. Human-like performance, which is prone to include immoral actions, may not be acceptable in machines, but moral perfection may be computationally unattainable. The risks posed by autonomous machines ignorantly or deliberately harming people and other sentient beings are great. The development of machines with enough intelligence to assess the effects of their actions on sentient beings and act accordingly may ultimately be the most important task faced by the designers of artificially intelligent automata.

Keywords: moral agency, computational ethics, moral Turing Test

1. Introduction

A good web server is a computer that efficiently serves up html code. A good chess program is one that wins chess games. There are some grey areas and fuzzy edges, of course. Is a good chess program one that wins most games or just some? Against just ordinary competitors or against world class players? But in one sense of the question, it is quite clear what it means to build a good computer or write a good program. A good one is one that fulfills the purpose we had in building it.

However, if you wanted to build a computer or write a program that is good in a moral sense, that is *a good moral agent*, it is much less clear what would count as success. Yet as artificial intelligence moves ever closer to the goal of producing fully autonomous agents, the question of how to design and implement an artificial moral agent becomes increasingly pressing. Robots possessing autonomous capacities to do things that are useful to humans will also have the capacity to do things that are harmful to humans and other sentient beings. How to curb these capacities for harm is a topic that is beginning to move from the realm of science fiction to the realm of

Journal of Experimental & Theoretical Artificial Intelligence ISSN 0952-813X print/ISSN 1362-3079 online © 2000 Taylor & Francis Ltd
http://www.tandf.co.uk/journals

real-world engineering problems. As Picard (1997) puts it: 'The greater the freedom of a machine, the more it will need moral standards'.

Attempts to build an artificial moral agent (henceforth AMA) are stymied by two areas of deep disagreement in ethical theory. One is at the level of moral principle: ethicists disagree deeply about what standards moral agents ought to follow. Some hold that the fundamental moral norm is the principle of utility, which defines right actions and policies in terms of maximizing aggregate good consequences, while others hold that certain kinds of actions are unjustifiable even if a particular token of the type would maximize aggregate good. The other level of disagreement is more conceptual or even ontological: apart from the question of what standards a moral agent ought to follow, what does it mean to *be* a moral agent? A suitably generic characterization might be that a moral agent is an individual who takes into consideration the interests of others rather than acting solely to advance his, her, or its (henceforth its) self-interest. But would a robot which had been programmed to follow certain standards be, *ipso facto*, a moral agent? Or would the robot have to also be capable of thinking about what it is doing in certain ways, for example by explicitly using certain norms in its decision procedure? Would it also have to be able to conceive of what it was doing as taking the moral point of view? More strongly still, if a moral agent must be autonomous in a rich sense, would it have to be capable of misapplying the standards in question, or even intentionally and knowingly disobeying them?

This paper surveys such perplexing problems facing attempts to create an AMA. The following section discusses in greater detail the two areas of disagreement in ethical theory just sketched. Subsequent sections turn to a consideration of the computational difficulties facing various approaches to designing and evaluating AMAs.

2. Moral agency and moral norms

Disagreement among ethical theorists about which norms moral agents ought to follow and disagreement about what it means to be a moral agent are interrelated. In this paper discussion is restricted to several of the best-known and most widely debated approaches to ethical theory. Even such a restricted survey is adequate to show the interrelations between disagreements about norms and about what it means to be a moral agent, for the connections become apparent as soon as two of the best-known approaches are sketched. The two approaches considered are utilitarianism, on the one hand, and Kant's use of the 'categorical imperative' on the other.

As a general school of thought, utilitarianism is the view that the best actions and institutions are those which produce the best aggregate consequences. Although there is disagreement even among utilitarians about which consequences matter in this estimation, the classical utilitarians (Bentham 1780, Mill 1861 and Sidgwick 1874) were sentientists, holding that effects on the consciousness of sentient beings are ultimately the only events of direct moral significance. Roughly, the classical utilitarians held that the best actions are those which produce the greatest happiness for the greatest number.

Mill famously held that 'it is better to be a human being dissatisfied than a pig satisfied', (1957 [1861], p.14) and he emphasized that one of the things which makes human happiness qualitatively superior to animals' (so that a relatively unhappy human might still be leading a better life than a thoroughly happy pig) was humans' capacity for moral agency. Mill held, further, that our evaluations of a moral agent's general character ought to be kept separate from our evaluation of any particular action by that agent:

> utilitarian moralists have gone beyond almost all others in affirming that the motive has nothing to do with the morality of the action, though much with the worth of the agent. He who saves a fellow creature from drowning does what is morally right, whether his motive be duty or the hope of being paid for his trouble (Mill 1957 [1861], pp.23–24).

There is, then, a clear sense of 'morally good' which a utilitarian can apply to an agent's actions irrespective of how the agent decided upon that action. Not so for Kant. According to Kant, for an action to be *morally* good, the action must, as he put it, be done out of respect for the categorical imperative. An enormous literature exists debating what exactly Kant meant by 'the categorical imperative' and by 'acting out of respect for it'. Here we follow one account of the categorical imperative and assume that 'acting out of respect for it' means simply that the agent acted as it did because it determined that the action in question was consistent with the categorical imperative. In this sense of the term, an action cannot be morally good unless the agent in fact reasoned in certain fairly complex ways.

Kant offers numerous formulations of the categorical imperative, but one of the most widely discussed is this: 'Act only on that maxim through which you can at the same time will that it should become a universal law' (1948 [1785], p.88). Singer (1971) provides a particularly insightful interpretation of this principle which we assume for purposes of this paper. By a maxim Kant means, roughly, an explicit and fully stated principle of practical reason. Specifically, a maxim has three elements: a goal which the agent proposes to achieve by acting on it; a means or course of action by which the agent proposes to achieve that goal; and a statement of the circumstances under which acting in that way will achieve the goal in question. The categorical imperative is a negative test, that is, it does not tell you which specific maxims to act on, rather, it requires you to never act on a maxim which you could not 'at the same time will that it should become a universal law'. By this we take him to mean (following Singer) that you could effectively achieve your goal in a world in which everyone sought to achieve the same goal by acting the same way in similar circumstances.

As Singer shows, this way of understanding the categorical imperative saves Kant from certain infamous objections. For present purposes, what is important, however, is that for Kant, a morally good action is one which is done because the agent has put the maxim of its action to the test of the categorical imperative and seen that it passes. In this way the motive of the action (trying to be good by sticking to what the categorical imperative requires) is essential to its being morally good, as are the specific deliberations involved in deciding that the maxim of one's action passes the categorical imperative.

In Kant and Mill, then, we find very different moral principles tied to very different conceptions of what a good moral agent is. In Mill's utilitarian terms, we might say that an agent is morally good to the extent that its behaviour positively affects the aggregate good of the moral community. In this sense a robot could be said to be a morally good agent to the extent that it has been programmed to act consistently with the principle of utility, regardless of how this behavioural result is achieved. For Kant, however, any claim than an agent is morally good (on either a specific occasion or in general) implies claims about the agent's internal deliberative processes. On Kant's view, to build a good AMA would require us to implement certain specific cognitive processes and to make these processes an integral part of the agent's decision-making procedure.

One further point about Kant illustrates another fundamental question about what would count as success in constructing an AMA. Kant held that the categorical imperative is only an 'imperative' for humans: 'for the *divine* will, and in general for a *holy* will, there are no imperatives: "*I ought*" is here out of place, because "*I will*" is already of itself necessarily in harmony with the law' (1948 [1785], p.81). If, as Kant appears to think, being a moral agent carries with it the need to *try* to be good, and thus the capacity for moral failure, then we will not have constructed a true artificial *moral* agent if we make it incapable of acting immorally. Some kind of autonomy, carrying with it the capacity for failure, may be essential to being a real *moral* agent. However, as we suggest below, the basic goals when constructing an *artificial* moral agent are likely to be very different than when raising a *natural* moral agent like a child. Accordingly, it may be acceptable to program a computer to be incapable of failure but unacceptable to attempt the analogue when raising a child.

3. A 'Moral Turing Test'?

In both ethical theory and day-to-day talk about ethics, people disagree about the morality of various actions. Kant claimed that it is always immoral to lie, no matter what the consequences (although Singer argues that Kant's own principles do not entail this conclusion). A utilitarian would deny it, holding instead that lying is justified whenever its consequences are sufficiently good in the aggregate. And day-to-day life is rife with disagreements about the morality of particular actions, of lifestyle choices and of social institutions.

In the face of such diverse views about what standards we ought to live by, an attractive criterion for success in constructing an AMA would be a variant of Turing's (1950) 'Imitation Game' (aka the Turing Test). In the standard version of the Turing Test, an 'interrogator' is charged with distinguishing a machine from a human based on interacting with both via printed language alone. A machine passes the Turing Test if, when paired with a human being, the 'interrogator' cannot identify the human at a level above chance. Turing's intention was to produce a behavioural test which bypasses disagreements about standards defining intelligence or successful acquisition of natural language. A Moral Turing Test (MTT) might similarly be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human 'interrogators' cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent.

One limitation on this approach is the emphasis it places on the machine's ability to *articulate* moral judgments. A Kantian might well be satisfied with this emphasis, since Kant required that a good moral agent not only act in particular ways but act as a result of reasoning through things in a certain way. Just as clearly, however, both a utilitarian approach and common sense suggest that the MTT places too much emphasis on the ability to articulate one's reasons for one's actions. As seen above, Mill allows that various actions are morally good independently of the agent's motivations, and many people think that young children, or even dogs, are moral agents even though they are incapable of articulating the reasons for their actions.

To shift the focus from conversational ability to action, an alternative MTT could be structured in such a way that the 'interrogator' is given pairs of descriptions of actual, morally-significant actions of a human and an AMA, purged of all references that would identify the agents. If the interrogator correctly identifies the machine at a level above chance, then the machine has failed the test. A problem for this version of the MTT is that distinguishability is the wrong criterion because the machine might be recognizable for acting in ways that are consistently *better* than a human in the same situation. So instead, the "interrogator" might be asked to assess whether one agent is less moral than the other. If the machine is not identified as the less moral member of the pair significantly more often than the human, then it has passed the test. This is called the 'comparative MTT' (cMTT).

There are several problems for the cMTT. First, one might argue that despite setting a slightly higher standard for machine behaviour than for humans, the standard is nevertheless too low. When setting out to design an artificial moral agent, we might think it appropriate to demand more than we expect of, say, human children. That is, the goal of AI researchers when constructing an AMA should not be just to construct a moral agent, but to construct an exemplary or even a perfect moral agent. The cMTT allows the machine's aggregate performance to contain actions that would be judged as morally wrong, and morally worse than the human actions, so long as on balance these do not cause the machine to be rated lower than the human. The cMTT could, in response, be tightened to require that the machine not be judged worse than the human in any of the pairwise comparisons of specific actions. But even if this restriction is added, there is a second way in which the standard might be thought too low, namely that the human behaviour is, itself, typically far from being morally ideal.

We humans typically muddle along making mistakes while harbouring private regrets about our moral lapses, which occur more frequently than perhaps we care to admit. But while we expect and, to a certain extent, tolerate human moral failures, it is less clear that we would, or should, design the capacity for such failures into our machines. Calculated decisions that result in harm to others are likely to be much less tolerated in a machine than in another human being. In other words, we shall probably expect more of our machines than we do of ourselves. And while murderous rampages are beyond the pale for both humans and AMAs, curbs on more mundane forms of immorality – the lying, cheating, and swindling of daily life – represent a much more difficult computational challenge, for these capacities are much more prominent in the 'grey' areas of morality, such as 'white' lies; the maxim 'never lie' has far more exceptions than 'never kill'.

If a standard is to be set for the behaviour of AMAs that is higher than the standard set for humans, where are such standards going to come from? As already indicated, computer scientists who turn to moral philosophy for an answer to this and related questions will find a field that does not provide a universally-accepted moral theory. Furthermore, as philosophical objectives are not exactly the same as those of engineers, there is a considerable gap between the available theories and the design of algorithms that might implement AMAs. Nevertheless, the philosophical investigations provide a framework for thinking about the implementation issues.

Two basic kinds of approach to this task are considered: (i) theoretical approaches that implement an explicit theory of evaluation (like the principle of utility or Kant's categorical imperative) thus providing a framework for the AMA to compute a morally preferred action and (ii) modelling approaches that either implement a theory of moral character (e.g. virtue theory) or that use learning or evolution to construct systems that act morally. Approaches in both categories present interesting computational challenges.

4. Theoretical approaches

Gips (1995) identifies two basic classes of theoretical approach to AMAs: 'consequentialist' theories and 'deontological' theories. Respectively, these theoretical approaches would attempt to implement either consequence-oriented reasoning (such as utilitarianism) or a rule- or duty-based form of reasoning (such as Kant's).

4.1. *Consequentialism*

The crucial problem for the consequentialist approach is that utilitarianism would seem to be a computational black hole. To implement the theory, the effects of an action on every member of the moral community must be assigned a numerical value. The sheer impracticality of doing this in real time for real world actions should be evident, especially if one considers the fact that the direct effects of every action ripple outwards to further effects that also affect aggregate utility. We are confident that interactions between these effects would make practical computation of long term consequences an intractable problem. Furthermore, if utilities must be computed for as long as an action has an effect in the world, potentially for all time, there is the risk of a non-terminating procedure. Even if the termination problem can be solved, it would also require the implementation of a comprehensive scientific theory to predict the nature of these long-range effects.

To insist on knowing every last consequence of an action sets a standard much higher than we expect of humans. We don't typically judge the morality of an action for its effect on sentient beings many years on the future. Indeed all judgements would have to be suspended if we did so. One might, then, try to restrict the computational problem by establishing an horizon beyond which assessment is not required. Unfortunately this is unlikely to be a simple matter. One might try to establish a temporal horizon by setting a time limit for agent responsibility. Or one might try to establish a physical or social horizon by stipulating a number of links in a causal or social chain beyond which an agent is not held responsible. But for any horizon one establishes as the point at which utilitarian calculations are stopped, one can imagine an agent deliberately initiating a process that will result in enormous pain and suffering at some point beyond the horizon. Yet we would surely judge this to be an immoral act.

Perhaps the situation is not as hopeless as we have made it sound if some standard AI techniques of tree searching can be applied. The chess computer Deep Blue (Campbell 1997) does not need to search the entire space of chess moves into the indefinite future to play a good game. Techniques for abbreviating the search are applied. The algorithm can be directed towards accomplishing intermediate goals, using approximate methods of evaluation, and working backwards from a desired outcome to determine a plan. Perhaps similar strategies could be adapted to moral decisions guided by utilitarian principles.

One way to deal with this problems might be to implement a hybrid system. Such a hybrid system might involve initially consequentialist computations out to a specified limit, at which point more abstract principles of duty (deontology) or character (virtue) come into play. Conversely one might implement a deontological system that can be overridden by consequentialist reasoning whenever the good consequences of an action 'clearly' outweigh the bad (Hare's (1981) two levels of moral reasoning). Such hybrid systems must, however, face the computational difficulties inherent in deontological approaches.

4.2. *Deontology*

'Deontology' is a term of art that refers to the notion of duty. According to deontological theories, actions are to be assessed according to their conformity with certain rules or principles, like Kant's categorical imperative. Other examples are Gert's (1988) system based on ten simple moral rules, Asimov's (1990) three laws of robotics, and the 'golden rule' (treat others as you would wish them to treat you).

From a computational perspective, a major problem with most deontological approaches (with the possible exception of Kant's) is that there is the possibility of conflict between the implied duties. Such conflicts are, of course, a major plot device in Asimov's fiction. Notably, it is even the case that a deadlock can result as a consequence of Asimov's first law alone. The law states that a robot may not injure a human being, or, through inaction, allow a human being to come to harm. A deadlock will result when a robot is faced with a choice between action and inaction where in either case the result will be the injury of one human being but no injury to another. The other two of Asimov's laws do not resolve the deadlock as they pertain only to following human orders and self-preservation. While this makes for interesting fiction, it is not a good basis for AMA design.

Other deontological systems, for example Gert's ten rules (1988), also contain rules that will inevitably conflict. As Gips (1995) reports, Gert's theoretical solution is to allow a rule to be disobeyed so long as an impartial rational observer could publicly advocate that it may be disobeyed. At this point we note that anyone interested in a computational implementation of this theory must wonder how an AMA is supposed to decide whether an impartial observer would allow such a rule violation without implementing another moral theory to make that determination. A further problem for deontological systems attempting to provide a list of rules to cover all situations is that such a list may fail to be comprehensive.

More abstract deontological theories, such as Kant's categorical imperative and the golden rule, attempt to avoid both the problem of comprehensiveness and the problem of conflicts between rules by formulating a single rule intended to cover all actions. However, they bring other computational problems because they are cast at such an abstract level. For instance, to determine whether or not a particular action satisfies the categorical imperative, it is necessary for the AMA to recognize the goal of its own action, as well as assess the effects of all other (including human) moral agents' trying to achieve the same goal by acting on the same maxim. This would require an AMA to be programmed with a robust conception of its own and others' psychology in order to be able to formulate their reasons for actions, and the capacity for modelling the population-level effects of acting on its maxim – a task that is likely to be several orders of magnitude more complex than weather forecasting (although it is quite possibly a task for which we humans have been equipped by evolution).

Similarly, to implement the golden rule, an AMA must have the ability to characterize its own preferences under various hypothetical scenarios involving the effects of others' actions upon itself. Further, even if it does not require the ability to empathize with others, it must at least have the ability to compute the affective consequences of its actions on others in order to determine whether or not its action is something that it would choose to have others do to itself. And it must do all this while taking into account differences in individual psychology which result in different preferences for different kinds of treatment.

5. Models of morality

'Theoretical' approaches that implement an explicit ethical theory in an AMA are computationally complex. An alternative approach to producing autonomous moral agents would be to bypass programming them to use explicit rules or principles in deciding what to do, and attempt, instead, to produce agents which happen to make morally proper decisions by modelling the moral behaviour of humans. Three approaches are discernible in this general category. The first, with an ancient history, is to consider a moral agent as one who possesses certain virtues which guide the agent's behavior and to attempt to program those virtues directly (see also Coleman 1999). A second approach involves the development of a model of moral agency by an associative learning process. The third approach consists in simulating the evolution of moral agents.

5.1. *Virtue approaches*

The basic idea underlying virtue ethics is that character is primary over deeds because good character produces good deeds. As Gips (1995) points out, virtue approaches frequently get assimilated to deontological approaches as the relevant virtues get translated into duties. Therefore, insofar as this assimilation is appropriate, virtue approaches raise exactly the same computational issues as deontological approaches, including the problem of dealing with conflicts between competing virtues, such as honesty and compassion.

Even if virtue approaches should not be assimilated to deontology, and can be considered as providing a separate model of moral character, the computational complexity of mapping relatively abstract character traits onto real actions looms as large as the problem of programming the computational use of rules for moral behaviour. Specifying, for example, that an AMA should have the character of honesty, requires an algorithm for determining whether any given action is honestly performed. But as we have learned from various of Plato's dialogues (Hamilton and Cairns 1961), it is no easy task to formulate definitions of such traits. It will be correspondingly difficult to determine whether particular actions conform to the virtues.

It is also difficult to come up with a comprehensive list of virtues that would cover all the scenarios that an AMA might find itself in. If available, such a list of virtues would provide a top-down specification for a model of moral agency. A more tractable approach, however, might be to develop models of moral agency from the bottom up. Each of the next two approaches fall in that category.

5.2. *Associative learning*

Just as young children learn appropriate behaviour from experience, one might approach the development of AMAs by way of a simulated childhood consisting of a training period involving feedback about the moral acceptability of actions. (This approach is endorsed by Dennett (1997) as part of the Cog robot project (Brooks *et al.* 1999)).

The implementation details, whether artificial neural networks or some other form of associative learning scheme, need not particularly concern us here as we know that the learning algorithms are computationally tractable. However, anyone taking this approach should be concerned about the quality of the feedback. A simple binary indication of the acceptability or unacceptability of an action may not be adequate for training purposes. The psychological literature on moral development seems to indicate that the best moral training involves embedding approval and disapproval in a context of reasons for those judgements (Damon 1999). As described by Damon, motivation by punishment and reward features at the lowest, self-interested stage of moral development represented in Kohlberg's scheme (Kohlberg 1984). At the second level comes social approval, divided into concern for the opinions of others and respect for social structures such as the law. Abstract ideals are reached only at the third level on Kohlberg's moral development chart, and empirical research on young males suggests that this abstract level is achieved by only a small fraction by their mid-twenties. (According to Damon, while early research suggested gender differences, subsequent and better-controlled studies have failed to support this finding.)

The point for computationalists is that the simplest associationist techniques based on binary-valued feedback are unlikely to produce fully satisfactory models of moral agency. If an artificial moral agent is to pass some variant of the cMTT, it is likely to require a capacity for abstract moral reasoning. While it must be possible ultimately to implement such reasoning using, for example, artificial neural networks, AI is a long way from understanding the network architecture that is required to do so. Nevertheless, it is our belief that the implementation of simpler schemes will be an important step towards the development of more sophisticated systems, even if the products of these simpler schemes are unlikely to pass the cMTT.

5.3. *Evolutionary/sociobiological approaches*

Another approach to modelling moral agency is through simulated evolution or artificial life. Combining sociobiology and game theory, Danielson (1992) develops a concept of 'functional' morality that is based on the idea that rationality is the only necessary quality that an agent must possess to be a moral agent (see also Harms *et al.* 1999 and Skyrms 1996). This approach is exemplified in an iterated Prisoner's Dilemma (PD) game (Axelrod and Hamilton 1981). By being iterated, no player has knowledge of when the game will end; it could continue throughout a lifetime, or involve only one PD interaction with a certain player. Examples of the PD abound in nature, and are evident when animals interact socially, such as: predatory alert signals, mutual grooming, sharing of food and group raising of young. Through evolution, organisms which have mutually iterated PD interactions evolve into a stable set of co-operative interactions. Most importantly for our present purposes, it has been shown that it is functionally optimal if an organism co-operates with other organisms, or, as Dawkins (1989) stated 'nice guys finish first'. The 'moral rules' that emerge from this process have no higher justification than survival values, for in the iterated game-theoretical scenarios it is purely because it is in the best interests of rational agents that they co-operate and behave in a way that gives the appearance of morality.

The evolutionary models developed by Danielson, Harms *et al.* and Skyrms show what can be achieved with computational techniques and further investigations should be encouraged. But, as Danielson admits, real-world morality has evolved in scenarios far more complex than the rather simplistic games that constitute the artificial environments for their evolutionary simulations. In scaling these environments to more realistic environments, evolutionary approaches are likely to be faced with some of the same shortcomings of the associative learning approaches: namely that sophisticated moral agents must also be capable of constructing an abstract, theoretical conception of morality. And while it is axiomatic that humans have evolved that capacity, the question of whether a simulated evolutionary approach is the most effective computational technique for developing AMAs is far from being settled.

6. The role of emotions: perfect theorizers, imperfect practitioners?

The discussion so far has omitted any mention of the role of emotion in actual human morality. Emotions undoubtedly provide important motivators for human behaviour. Perfect knowledge of a moral theory does not guarantee action in conformity with that theory. Thus one might well imagine an agent who is fully capable of deciding whether certain actions are moral while being completely unmotivated to act morally. Such an agent might frequently act in a way he or she knew to be wrong. Indeed this is the description of some sociopaths.

As well as providing motivation, it has also been argued that emotions are essential to intelligence generally (see Picard 1997 for review). It has also been argued that emotions provide moral knowledge: for instance, the feeling of shame that accompanies the memory of some action might serve to let you know that the action was morally wrong. Whether or not this epistemological claim is true, human practical morality is undoubtedly driven by a complex blend of reason and emotion. Among those emotions, empathy for others plays a strong role in determining actual moral behaviour. Where empathy for others is lacking, morally unacceptable behaviour often results (Damon 1999). Emotions are also important for evaluating alternative futures but emotion is also a double-edged sword. While emotional engagement and empathy are requirements for human practical morality, the very existence of passionate emotions is also what, in many cases, causes us to do things that, in the cold light of day, or from a neutral perspective, we judge immoral.

If emotions are essential for intelligence or moral agency, computer scientists undoubtedly have a long way to go before building an AMA. But perhaps emotional engagement is not essential for AMAs even if it is required for human practical morality. In support of this view we note that emotion does not seem to be a requirement for autonomous behaviour *per se*. Robots designed to navigate through a building, seeking soda cans, are not motivated by any emotions to do so. Their systems assess alternative pathways and select among them. The choices made by these robots are autonomous: no external guidance is needed for them to select one pathway rather than another. Accordingly, Deep Blue has no emotional engagement in the game of chess, yet the choices it makes about which pieces to move where are, nonetheless, autonomously made. This very lack of passion contributes, arguably, to making Deep Blue a better chess player, or, at least, a more reliable player, than any human. Rattle the world champion's nerves and his chess game goes to pieces; Deep Blue marches on relentlessly. By analogy, then, one might even hope for *better* moral performance from a machine that cannot be distracted by sudden lusts, than one expects from a typical human being.

7. Conclusions

This is an exciting area where much work, both theoretical and computational, remains to be done. Top-down, theoretical approaches are 'safer' in that they promise to provide an idealistic standard to govern the actions of AMAs. But there is no consensus about the right moral theory, and the computational complexity involved in implementing any one of these standards may make the approach infeasible. Virtue theory, a top-down modelling approach, suffers from the same kinds of problems. Bottom-up modelling approaches initially seem computationally more tractable. But this kind of modelling inherently produces agents that are liable to make mistakes. Also, as more sophisticated moral behaviour is required, it may not be possible to avoid the need for more explicit representations of moral theory. It is possible that a hybrid approach might be able to combine the best of each, but the problem of how to mesh different approaches requires further analysis.

We think that the ultimate objective of building an AMA should be to build a morally praiseworthy agent. Systems that lack the capacity for knowledge of the effects of their actions cannot be morally praised or blamed for effects of their actions (although they may be blamed in the same sense as faulty toasters may be blamed for the fires they cause). Deep Blue is blameless in this way, knowing nothing of the consequences of beating the world champion, and therefore not morally responsible if its play provokes a psychological crisis in its human opponent. Essential to building a morally praiseworthy agent is the task of giving it enough intelligence to assess the effects of its actions on sentient beings, and to use those assessments to make appropriate choices. The capacity for making such assessments is critical if sophisticated autonomous agents are to be prevented from ignorantly harming people and other sentient beings. It may be the most important task faced by developers of artificially intelligent automata.

References

Asimov, I, 1990, *Robot Visions* (New York: Penguin Books).

Axelrod, R. and Hamilton, W., 1981, The evolution of cooperation. *Science*, **211**, 1390–1396.

Bentham, J., 1907 [1780], *An Introduction to the Principles of Morals and Legislation* (Oxford: Clarendon Press).

Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M., 1999, The cog project: building a humanoid robot. In C. L. Nehaniu (ed.) *Computationfor Metaphors, Analogy and Agents. Lecture notes in Artificial Intelligence*, vol. 1562 (Berlin: Springer-Verlag), pp. 52–87. Prepublication version at http://www.ai.mit.edu/projects/cog/Publications/CMAA-group.pdf

Campbell, M., 1997, An enjoyable game: how HAL plays chess. In D. Stork (ed.) *HAL's Legacy: 2001's Computer as Dream and Reality* (Cambridge, MA: MIT Press), pp. 75–98.

Coleman, K., 1999, Android Arete: virtue ethics for computational agents. Department of Philosophy, University of British Columbia, Canada. Paper presented at the 14th Annual Conference on Computing and Philosophy, Carnegie Mellon University, Pittsburgh, PA, August 5–7.

Damon, W., 1999, The moral development of children. *Scientific American*, **281**(2), 72–78.

Danielson, P., 1992, *Artificial Morality*: *Virtuous Robots for Virtual Games* (New York: Routledge).

Dawkins, R., 1989, *The Selfish Gene* (New York: Oxford University Press).

Dennett, D., 1997, When HAL kills, who's to blame? In D. Stork (ed.), *HAL's Legacy: 2001's Computer as Dream and Reality* (Cambridge, MA: MIT Press), pp. 351–365.

Gert, B., 1988, *Morality* (New York: Oxford University Press).

Gips, J., 1995, Towards the ethical robot. In K. Ford, C. Glymour and P. Hayes (eds), *Android Epistemology* (Cambridge. MA: MIT Press), pp. 243–252.

Hamilton, E., and Cairns, H., 1961, *The Collected Dialogues of Plato, Including the Letters* (Princeton, NJ: Princeton University Press).

Hare, R., 1981, *Moral Thinking: Its Levels, Methods, and Point* (New York: Oxford University Press).

Harms, W., Danielson, P., and MacDonald, C., 1999, Evolving artificial moral ecologies. The Centre for Applied Ethics, University of British Columbia, available online: http://eame.ethics.ubc.ca/eameweb/eamedesc.html

Kant, I., 1948 [1785], *Groundwork of the Metaphysic of Morals*, translated by H. J. Paton (New York: Harper Torchbooks).

Kohlberg, L., 1984, *The Psychology of Moral Development.* (San Francisco: Harper and Row).

Mill, J. S., 1957 [1861], *Utilitarianism* (Indianapolis: The Liberal Arts Press).

Picard, R., 1997, *Affective Computing* (Cambridge, MA: MIT Press).

Sidgwick, H., 1874, *The Methods of Ethics* (London: MacMillan).

Singer, M. G., 1971, *Generalization in Ethics* (New York: Atheneum Press).

Skyrms, B., 1996, *Evolution of the Social Contract* (London: Cambridge University Press).

Turing, A., 1950, Computing machinery and intelligence. *Mind*, **59**, 433–460.