

Calculated morality: ethical computing in the limit

Colin Allen

Department of Philosophy, Texas A&M University, College Station, Texas
77843-4237, USA. <colin-allen@tamu.edu>

Abstract

When a person is described as “too calculating” this generally means that he lacks a certain degree of ethical sensitivity, having a tendency to deliberately exploit the altruism of others in order to further his own well-being. Computers are the ultimate calculating devices, and one might reasonably worry that the calculated nature of decisions made by any artificially intelligent system will result in behavior that is not recognizably moral. There are two ways to try to surmount this problem. One is to try to write high level rules that could be used to filter out problematic behaviors. Kant’s categorical imperative is an example in this category. The other approach is to abandon the search for explicitly rule-based approaches and seek other means of guiding behavior. For example, providing artificial moral agents with human-like emotions might serve to keep certain anti-social tendencies in check. Each approach has strengths, but both are deeply problematic. Consequently, we should expect that our artificial moral agents will be as subject to making ethical mistakes as the next person.

Keywords: computer, artificial intelligence, morals, ethics

Approaches to the Design of Artificial Moral Agents

The development of increasingly autonomous software agents and robotic systems forces us to take seriously the need for such agents to have a capacity for moral reasoning. Autonomous systems must make choices in the course of flexibly fulfilling their missions, and some of those choices will have potentially harmful consequences for humans, and other subjects of moral concern. An autonomous system that ignorantly causes harm might not be morally blameworthy, any more than a toaster that catches fire can itself be blamed (although its designers may be at fault). But, in complex automata, this kind of blamelessness provides insufficient protection for those who may be harmed. If an autonomous system is to minimize harm, it must be cognizant of possible harmful consequences of its actions and it must select its actions in light of this knowledge. Such a system would be an artificial moral agent and the proper design of such systems is perhaps the most important and challenging task facing developers of fully autonomous systems (Allen et al., 2000). Picard (1997) puts it well when she writes, “The greater the freedom of a machine, the more it will need moral standards.”

There are two basic kinds of strategy that designers of artificial moral agents may pursue, although these may be combined in hybrid strategies (Allen et al., 2000). The first strategy is theoretical, or “top-down”, as it would attempt to regulate the behavior of artificial moral agents via algorithmic implementation of a specific moral theory represented as a set of rules. The other

strategy is a modelling, or “bottom-up”, strategy that attempts to develop artificial moral agents by creating contexts in which moral behavior can be learned, evolved, or otherwise acquired without the need for an explicitly coded set of moral rules. Both top-down, theoretical approaches to ethics, such as deontological (duty-based) approaches or consequentialist (outcomes-based) approaches, and bottom-up, learning approaches which take human behavior as a model, or evolutionary approaches which take social conditions as a source of selection pressure for ethical behavior, present enormous challenges to the designers of artificial moral agents.

Even if the choice of correct moral theory to be implemented can be settled, top-down approaches face significant issues of computational tractability. The Kantian deontologist, for instance, is required to determine whether or not his motive for action could be universally willed without contradiction, but the determination of this fact requires complicated reasoning about the logical and empirical consequences of everyone acting in accord with a given principle. (And this itself assumes that we have settled upon an uncontroversial reading of Kant’s (1785) categorical imperative, which is far from being a given -- for more details in the context of artificial moral agents, see Allen et al., 2000.) The Utilitarian moral agent must deliberate by computing the contribution of her actions to the aggregate good according to the rule that the best action is the one with the most positive effect on utility. This task requires detailed empirical foreknowledge of the effects of her actions and the ability to reason about many alternative scenarios. No imaginable system can compute the consequences of an action on aggregate happiness for the rest of time, so an unlimited implementation of utilitarian theory would seem to be beyond reach. Likewise, the capacity to consider all the possibly relevant logical and empirical consequences of any given maxim of action would seem to be beyond the reach of any imaginable artificial system. Thus there are limits to rule-based computational ethics. Nevertheless, it is reasonable to suppose that artificial moral agents, through the wonders of silicon hardware, could take such calculations further than human brains can go. The consequences of this fact for artificial morality are discussed in the final section.

In contrast to top-down approaches, bottom-up approaches might attempt to construct functioning models of moral decision making in artificial systems either by exposure to examples of morally praiseworthy behavior during a learning or training phase, or perhaps by simulating the evolution of moral agents. A major advantage of such modelling approaches is that they might be developed without resolving difficult philosophical issues about which moral theory is the correct theory. By taking agreed upon examples of moral behavior as the training set, or by simulating the social conditions in which morality has evolved, autonomous agents might, if we are sufficiently clever in designing them to learn or evolve, acquire a set of dispositions to act morally. But bottom-up approaches to artificial moral agents are unlikely to inspire public confidence because these methods seem *prima facie* unlikely to produce artificial moral agents that are any less prone to engage in immoral actions than humans. Neither learning nor evolutionary processes have proven to be entirely reliable methods for producing morally good human beings. We currently understand too little about the role of education and observational learning in developing moral character, and even less about the conditions leading to the evolution of moral agents, so it’s not presently clear why one should expect these techniques to lead reliably to the development of robust artificial moral agents.

We tolerate a certain amount of moral failing among our fellow humans. It is less clear that we would tolerate the same degree of failings in artificial systems. Is this a double standard? Perhaps so, but that's a topic for another time. The point here is just that, despite their limitations, top-down approaches seem to offer the better prospect for superior moral performance because they offer prospect of governing behavior by consistent rules that can serve to filter out problematic behaviors. Supporting the idea that rule-based approaches to morality will prove fundamental to the design of artificial moral agents, there is the tendency for bottom-up approaches to gravitate towards hybrid approaches that require explicit attention to ethical theory. This is because fully competent human moral agents incorporate theoretical elements into their decisions, so any model of such an agent requires some capacity for reasoning theoretically about morality. The human species has, after all, evolved a capacity for moral theorizing, and that capacity develops individually (albeit to varying degrees) in each of us as members of the species (see Damon 1999).

Emotionless Morality?

How is moral performance in artificial moral agents to be evaluated? Allen et al. discussed ways of formulating a moral Turing test in which, like the test on which it is modeled (Turing, 1950), a "blind" observer is asked to compare the behavior of a machine to human behavior. Unlike the Turing test, which uses statistical indistinguishability of human and machine responses as the criterion for machine success, the comparative version of the moral turing test proposed by Allen et al. requires only that the machine should not be judged less moral on average. This version of the test allows the possibility that the machine could be distinguishable from the human by being consistently judged more moral.

If moral deliberation requires sophisticated computation of logical or actual consequences of actions, one would expect that the superior computational abilities would be correlated with superior moral capacities. This connection between superior reasoning ability and morality has been explored in science fiction: for instance, in *Star Trek's* Mr. Spock and Commander Data who are often portrayed as acting more selflessly and in a more principled fashion than their human counterparts. And yet there is a strong countercurrent in the same genre which questions whether real moral competency is possible when emotions are suppressed or wholly absent. The more impetuous choices of emotional humans are often portrayed as more authentically moral than the more calculated responses of nonhuman super-reasoners.

Perhaps this insistence on emotional engagement is nothing more than human narcissism. We can admit that emotional empathy is a prime motivator for human moral actions, without it following that morality is impossible in the absence of emotions. We can also admit that human moral choices have been shaped in positive ways by the presence of emotions, while being forced to concede that emotions frequently lead to immoral actions by impeding careful moral deliberation. There can be no simple relationship however between emotions and moral behavior, for the very same emotions (e.g. love or pride) can sometimes lead to moral behavior, and sometimes to immoral behavior.

But let us grant that the emotions are an essential part of the story about of *human* morality, the question remains whether emotions are essential to all moral agents, and in

particular whether superior moral performance in artificial moral agents is more likely in the absence of some or all emotions (as suggested by Allen et al., 2000). Even if the capacity to experience emotions directly is not required in order to produce human-like or super-human moral performance, and even if the absence of emotions actually causes artificial moral agents to be superior moral agents, such agents will most likely will require the capacity to understand and reason about the emotions of others (Picard, 1997) for without such understanding it would not be possible to assess adequately the effects of an action upon others. For the Utilitarian, the importance of understanding them is straightforwardly obvious: both positive and negative emotions figure into the utilitarian calculus of goods and harms and so must be considered among the consequences of any action. The Kantian case is (as usual!) more complicated, in that the relationship between one's duties to others and the emotions that one's actions cause in others is less straightforward. Nonetheless it is hard to see how any competent Kantian agent could ignore the emotions in determining whether or not another human being is being treated as an end in herself or as a means to some end.

It is easy to lose sight of the importance of considering emotional effects when considering the kinds of life-and-death ethical dilemmas to which both television scripts and undergraduate courses in ethics tend to gravitate. With lives in the balance, emotional responses can seem inconsequential. But in fact, the bulk of human moral behavior is much more concerned with mundane matters of small favors and minor harms than it is with life for death. If I lie to you, it is much more likely that I manipulate your emotions by doing so than that I am lying about something that would kill you. If I help you by carrying a heavy suitcase, it is unlikely that I will have saved your life, although I may make you happy. These mundane items do not feature in grand discussions of morality, and yet, for the design of artificial moral agents, I believe that they loom quite large.

Humans typically help others, including complete strangers, by reflexively performing relatively minor tasks, such as giving directions or holding open a door. These kinds of actions provide relatively minor benefits to those helped at relatively small cost to ourselves. Yet such actions are not typically motivated by recognition of a theoretical duty or by an explicit cost-benefit analysis according to a theory. Indeed, anyone who explains a decision to hold open a door or to give money to a charity on the grounds that it was required by a duty, or that it was calculated to increase aggregate utility, would typically be regarded as a bit odd, and perhaps less morally praiseworthy than one who responded that it just felt like the right thing to do. In an important sense, the more theoretical approaches are "too calculating" (although not, of course, in the sense that this phrase, in English, is often used to indicate excessive promotion of self interest at the expense of others). Emotional responses to those we encounter appear to play a significant role in guiding ethical behavior. The feeling that one should help another is an emotional response to a situation that may, in fact, run against the calculations of expected utility or other duties. Nonetheless, any agent that failed to act in such ways would be regarded as missing a key component of human morality. Furthermore, to return to the kinds of life and death issues that are the staple of introductory courses in ethics, we judge as attempts to rescue others at great personal risk to the self as morally praiseworthy, even when no duty requires such action, or when the odds of failure mean that a strict utilitarian calculation would indicated an expected reduction in expected utility from the action. The importance of empathy-driven responses to human moral

behavior suggests that theory-based approaches to the design of AMAs cannot completely ignore emotions.

Living with Moral Calculators

Whether the outcome of a top-down development process or the result of a bottom-up process leading to moral patterns of behavior, artificial moral agents are likely to be instantiated in systems that have fundamentally different, and in many ways more powerful computational capacities than humans. Imagine, if you will, artificial moral agents that share our values but that can see further into the future than we can, or that can draw out the logical consequences of their moral rules much further than we can. Such systems would be in a position to point out what, from the perspective of the implemented theory, are actual or potential moral mistakes. All of us know people (and some of us perhaps even are people) who would prefer to rely on the superior computational power of artificial moral agents to help us see the logical and empirical consequences of our actions rather than relying on more intuitive, less systematic approaches to moral decision making. Some people will no doubt find this to be a terrifying prospect.

In either case, the effects of interacting with artificial moral agents are likely to be profound. If the behavior or advice that results from an implementation of a particular ethical theory doesn't strike us as morally praiseworthy, we must choose between revising the theory and altering our conception of morality. In either case, by forcing us to confront the computational limits of our best moral theories, interactions with artificial moral agents have the potential to fundamentally alter our very sense of morality itself.

Acknowledgement

I wish to thank Gary Varner and Jason Zinser for helping me to develop the ideas that went into our (2000) paper.

References

Allen, C., Varner, G. and Zinser, J. (2000); A Prolegomena to Any Future Artificial Moral Agent: *Journal of Experimental and Theoretical Artificial Intelligence*. Vol. No. (pp. 251-261).

Damon, W. (1999); The Moral Development of Children: *Scientific American*. Vol. 281 No. 9 (pp. 72-78).

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

Picard, R. (1997); *Affective Computing*. MIT Press.

Turing, A. (1950); Computing Machinery and Intelligence. *Mind*. Vol. 59 (pp. 433-460).