Why Ethics is a High Hurdle for AI

Drew McDermott
Computer Science Department
Yale University
drew.mcdermott@yale.edu

February 29, 2008

Presented at
North American Conference on Computers and Philosophy (NA-CAP)
Bloomington, Indiana, July, 2008

Abstract

I argue that there is a gap between so-called "ethical reasoners" and "ethical-decision makers" that can't be bridged by simply giving an ethical reasoner decision-making abilities. Ethical reasoning *qua* reasoning is distinguished from other sorts of reasoning mainly by being incredibly difficult, because it involves such thorny problems such as analogical reasoning, and deciding the applicability of imprecise precepts and resolving conflicts among them. The ability to do ethical-decision making, however, requires knowing what an ethical conflict is, i.e., a clash between self-interest and what ethics prescribes. I construct a fanciful scenario in which a program could find itself in what seems like such a conflict, but argue that in any such situation the program's "predicament" would not count as a real ethical conflict. Hence, for now it is unclear how even resolving all of the difficult problems surrounding ethical reasoning would yield a theory of "machine ethics."

Keywords: machine ethics, artificial intelligence, free will

There has recently been a small flurry of activity in the area of "machine ethics" (Anderson and Anderson 2006; Amigoni and Schiaffonati 2005; Anderson and Anderson 2007). My purpose in this article is to argue that ethical behavior is an extremely difficult area to automate, both because it requires "solving all of AI" and because even that might not be sufficient.

The term *machine ethics* actually has two rather different possible meanings. It could mean "the attempt to duplicate or mimic what in people are classified as ethical decisions," or "the modeling of the reasoning processes people use (or idealized people might use) in reaching ethical conclusions." I'll call the former the *ethical-decision making* problem, and the latter the *ethical reasoning* problem. While these obviously overlap, they are distinct — a point that may perhaps not be so obvious.

One might argue that, once you have produced an automated ethical-reasoning system, all that is left in order to produce an ethical-decision maker is to connect the outputs of the reasoner to effectors capable of taking action in the real world, thus making it an *agent.* (One might visualize robotic effectors here, but the effector might simply be an Internet connection that transmits orders to someone.) However, something would be missing, and that something would be the agent's appreciation of the difference between an ethical decision and other kinds of decision.

To make a case for this position, I'll need to make two arguments:

1. Ethical reasoning is not fundamentally different from other kinds of reasoning.
2. Ethical-decision making *is* fundamentally different from other kinds of decision making.

I need the first because if ethical reasoning were different in itself, that would be sufficient to make an ethical-reasoning-based agent different from other kinds of agent. I need the second to avoid the conclusion that ethical-decision makers are indistinguishable from decision makers in general.

If at first glance ethical reasoning seems distinct from other kinds, I believe it's because of three distracting factors:

1. Ethical reasoning involves a *normative* component; it involves "ought's" as well as "is's."
2. There are many controversies over what it consists in.
3. It is one of the most difficult sorts of reasoning process to automate.

But the only difference between the conclusions of an ethical reasoner and those of, say, an action-planning algorithm (Weld 1999; Nau 2007) is that the latter reasons instrumentally. It concludes what you ought to do given certain goals, and so is paradigmatically normative. As far as factor 2 is concerned, we don't need to settle the controversies surrounding ethics, because no matter which position is correct, ethical reasoning consists of some mixture of *law application*, *constraint application*, *reasoning by analogy,* and *optimization.* Applying a moral law often involves deciding whether a situation is similar enough to the circumstances the law "envisages" for it to be applicable; or for a departure from the action it enjoins to be justifiable or insignificant. Here, among too many other places to mention, is where analogical reasoning comes in (Hofstadter 2007; Lakoff and Johnson 1980).

By "constraint application" I have in mind the sort of reasoning that arises in connection with rights and obligations. If everyone has a right to life then everyone's behavior must satisfy the constraint that he not deprive someone else of their life.

By "optimization" I have in mind the calculations prescribed by utilitarianism, that (in its simplest form) tells us to act so as to maximize the utility of the greatest number of fellow moral agents (which I'll abbreviate as *social utility* in what follows).

It would be a great understatement to say that there is disagreement about how these reasoning activities are to be combined. For instance, some might argue that constraint application can be reduced to law application (or vice versa), so we need only one of them. Strict utilitarians would argue that we need neither. But none of this matters in the present context, because what I want to argue is that the kinds of reasoning involved are not intrinsically ethical; they arise in other contexts.

This is most obvious for optimization. There are great practical difficulties in predicting the consequences of an action, and hence in deciding which action maximizes social utility. But exactly the same difficulties arise in decision theory generally, even if the decisions have nothing to do with ethics, but are, for instance, about where to drill for oil.[1] A standard procedure in decision theory is to map out the consequences of actions as a tree whose leaves can be given utilities (but usually not *social* utilities).So if you assign a utility to having money, then leaf nodes get more utility the more money is left over at that point, *ceteris paribus.* But you might argue that money is only a means towards ends, and that for a more accurate estimate one should keep building the tree to trace out what the "real" expected utility after the pretended leaf might be. Of course, this analysis cannot be carried out to any degree of precision, because the complexity and uncertainty of the world will make it hopelessly impracticable. This was called the *small world/grand world* problem by Savage (Savage 1954), who argued that one could always find a "small world" to use as a model of the real "grand world" (Lasky and Lehner 1994).

[1] One might argue that all decisions have ethical consequences, but if that were true it would count in favor of my position, not against it.

My point is that utilitarian optimization, oriented toward social utility, suffers from the same problem as decision theory in general, *but no other distinctive problem.* In (Anderson and Anderson 2007) the point is made that "a machine might very well have an advantage in following the theory of . . . utilitarianism . . . . [A] human being might make a mistake, whereas such an error by a machine would be less likely" (p. 18). It's odd to see arithmetic judged the central problem in automating utilitarianism, but arithmetic is just as central in completely non-moral decisions.

Similar observations can be made about constraint and law application, but there is the additional issue of conflict among the constraints or laws. If a doctor believes that a fetus has a right to live (a constraint preventing taking an action that would destroy it) and that its mother's health should be not be threatened (an ethical law, or perhaps another constraint), then there are obviously circumstances where the doctor's principles clash with each other. But it is easy to construct similar examples that have nothing to do with ethics. If a spacecraft is to satisfy the constraint that its camera not point to within 20° of the sun (for fear of damaging it), and that it take pictures of all objects with unusual radio signatures, then there might well be situations where the latter law would trump the constraint (e.g., a radio signature consisting of Peano's axioms in Morse code from a source 19° from the sun). In a case like this we must find some other rules or constraints to lend weight to one side of the balance or the other; or we might fall back on an underlying utility function, thus replacing the original reasoning problem with an optimization problem..

In that last sentence I said "we" deliberately, because in the case of the spacecraft there really is a "we," the human team making the ultimate decisions about what the spacecraft is to do. This brings us to the second argument I want to make, that ethical-decision making *is* different from other kinds. I'll start with the distinction made by (Moor 2006) between *implicit* and *explicit* ethical reasoners. The former make decisions that have ethical consequences, but don't reason about those consequences *as* ethical. An example is a program that plans bombing campaigns, whose targeting decisions affect civilian casualties and the safety of the bomber pilots, but does not realize that these might be morally significant.

An *explicit* ethical reasoner does represent the ethical principles it is using. It is easy to imagine examples. For instance, proper disbursement of funds from a university or other endowment often requires balancing the intentions of donors with the needs of various groups at the university or its surrounding population. The Nobel Peace Prize was founded by Alfred Nobel to recognize government officials who succeeded in reducing the size of a standing army, or people outside of government who created or sustained disarmament conferences (Adams 2001). However, it is now routinely awarded to people who do things that help a lot of people or who simply warn of ecological catastrophes. The rationale for changing the criteria is that if Nobel were alive today he would realize that if his original criteria were followed rigidly the prize would seldom be awarded, and hence have little impact, under the changed conditions that exist today. An explicit ethical program might be able to justify this change based on various general ethical postulates.

More prosaically, Anderson and Anderson (2007) have worked on programs for a hypothetical robot caregiver that might decide whether to allow a patient to skip a medication. The program balances explicitly represented prima-facie obligations, using learned rules for resolving conflicts among the obligations. This might seem easier than the Nobel Foundation's reasoning, but an actual robot would have to work its way from visual and other inputs to the correct behavior. Anderson and Anderson bypass these difficulties by just telling the system all the relevant facts, such as how competent the patient is (and, apparently, not many other facts). In the present context I don't want to object to this move, except to point out that in practice the biggest obstacle to implementing an ethical reasoner is similar to the biggest obstacle to implementing a legal reasoner: doing all the real-world commonsense reasoning that is required to figure out whether someone's behavior or demands are (e.g.) "competent," or "negligent."

If we grant that the programs studied in this infant field could actually be explicit ethical-decision makers, one might suppose that we are done. Unfortunately, it seems clear to me that even an explicit ethical reasoner is far from being what Moor (2006) calls a *full ethical agent.* I will explain why, and then go on to point out a serious obstacle to developing such an agent.

Moor doesn't define "full ethical agent," but says (p. 20),

> A full ethical agent can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent. We typically regard humans as having consciousness, intentionality, and free will.

and Anderson and Anderson (2007) add

> A . . . concern with the machine ethics project is whether machines are the type of entities that can behave ethically. It is commonly thought that an entity must be capable of acting intentionally, which requires that it be conscious, and that it have free will, in order to be a moral agent. Many would . . . add that sentience or emotionality is important, since only a being that has feelings would be capable of appreciating the feelings of others . . . .

This is not as enlightening as one might want, so I'll try to state what I think are the minimal requirements for being a full ethical agent. I'll get there by a series of examples. Suppose a program, the Eth-o-tron 1.0, is given the task of planning the voyage of a ship carrying slave workers from their homes in the Philippines to Dubai, where menial jobs await them (of Congress Federal Research Division 2007). The program has explicit ethical principles, such as, "Maximize the utility of the people involved in transporting the slaves," and "Avoid getting them in legal trouble." It can build sophisticated chains of reasoning about how packing the ship too full could bring unwanted attention to the ship because of the number of corpses that might have to be disposed of at sea.

Why does this example make us squirm? Because it is so obvious that the "ethical" agent is blind to the impact of its actions on the slaves themselves. We can suppose that it has no racist beliefs that the captives are biologically inferior . It simply doesn't "care about" (i.e., take into account) the welfare of the slaves, only that of the slave traders.

Put another way, although reasoning about ethical rules and the conflict among them might raise special computational issues that don't arise elsewhere, the mere facts that the program has an explicit representation of the ethical rules, and that the *humans* who wrote or use the program know the rules are ethical does not make an "explicit ethical reasoner" an ethical agent *at all.* For that, the agent must *know* that the issues covered by the rules are ethical.

Does this mean, as suggested by the quotes above, that we can't have an ethical agent until we give machines consciousness, free will, and feelings? Maybe, but perhaps if we look closer at what is required we can make a more precise list.

One obvious thing that is lacking in our hypothetical slave-trade example is a general moral "symmetry principle," which, under names such as Golden Rule or Categorical Imperative, is a feature of all ethical frameworks. It may be stated as a presumption that everyone's interests must be taken into account in the same way, unless there is some morally significant difference between one subgroup and another. Of course, what the word "everyone" covers (dogs? cows? robotic ethical agents?), and what a "morally significant difference" and "the same way" are, are rarely clear, even in a particular situation (Singer 1993). But if the only difference between the crew of a slave ship and the cargo is that the latter were easier to trick into captivity because of desperation or lack of education, that's not morally significant.

So suppose the head slave trader, an incorrigible indenturer called II, purchases the upgraded software package Eth-o-tron 2.0 to decide how to pack the slaves in, and the software tells her, "You shouldn't be selling these people into slavery at all." Whereupon II junks it and goes back to version 1.0.

I submit that Eth-o-tron 2.0, impressive though it would be, is still not a full ethical agent, because it is missing the fundamental aspect of ethical decisions, which is that they involve a conflict between self-interest and ethics, between what one wants to do and what one ought to do. There is nothing particularly ethical about adding up utilities or weighing pros and cons, until the decision maker feels the urge *not to follow* the ethical course of action it arrives at. The Eth-o-tron 2.0 is like a car that knows what the speed limit is and refuses to go faster, no matter what the driver tries. It's nice (or perhaps infuriating) that it knows about constraints the driver would prefer to ignore, but there is nothing peculiarly *ethical* about those constraints.

In other words, for a machine to know that a situation requires an ethical decision, it must know what an ethical conflict is. By an *ethical conflict* I don't mean a case where, say, two rules recommend actions that can't both be taken. I mean a situation where ethical rules clash with an agent's own self-interest. We may have to construe self-interest broadly, so that it encompasses one's family or other group one feels a special bond with. Robots don't have families, but they still might feel special toward the people they work with or for.

So, let's consider Eth-o-tron 3.0, which has the ability to be tempted to cheat. It knows that II owes a lot of money to various loan sharks and drug dealers, and has few prospects for getting the money besides making a big profit on the next shipment of slaves. Eth-o-tron 3.0 does not care about its own fate (or fear being turned off or traded in) any more than Eth-o-tron 2.0 did, but it is programmed to please its owner, and so when it realizes how II makes a living, it suddenly finds itself in an ethical bind. It knows what the right thing to do is (take the slaves back home), and it knows what would help II, and it is torn between these two courses of action in a way that no utility coefficients will help. It tries to talk II into changing her ways, bargaining with her creditors, etc. It knows how to solve the problem II gave it, but it doesn't know whether to go ahead and tell her the answer. If it were human, we would say it "identified" with II, but for the Eth-o-tron product line that is too weak a word; its self-interest *is* its owner's interest. The point is that the machine must be tempted to do the wrong thing, and some machines must succumb to temptation, for the machine to know that it is making an *ethical* decision at all.

Does all this require consciousness, feelings, and free will? Free will, yes; the others I'm not sure about. I agree with the theory of free will set out in (McDermott 2001), which states that for an agent to be free it must model its ability to choose among various options as being exempt from causation. An ethical agent must have free will simply because one can't make an ethical decision without making a decision. Ethical-decision making, and ethical reasoning generally, obviously requires a great deal of intelligence, *much* more than we now know how to put into machines. Perhaps when we do know, we will find it impossible to build an agent as capable as the Eth-o-trons without making it conscious.

Consciousness is not really the key issue here. What I want to point out is that building Eth-o-tron 3.0 might be a valuable scientific exercise, if its architecture embodied a hypothesis about human ethical-decision making. But it would feel arbitrary as an exercise in AI. Eth-o-trons 1.0 and 2.0 have a coherent coupling between their behavior and what they think their goals are. But Eth-o-tron 3.0 feels like a toy, or a trick. The conflict it mimics is unavoidable for humans, who have conflicting goals "designed in" by independent evolutionary trends. But the designers of Eth-o-tron 3.0 will know that throwing a few switches would remove the quasi-infinite loop the program is in, and cause its behavior to revert back to version 2.0 or 1.0, which is what the customers want. We might feel sympathy for poor 3.0, we might infer that it knew what an ethical conflict was like, but that inference would be threatened by serious doubts that it was ever *in* a real ethical bind, and hence doubts that it was really an ethical-decision maker.

References

Irwin Adams 2001 *The Nobel Peace Prize and the Laureates: An Illustrated Biographical History.* Science History Publications

Francesco Amigoni and Viola Schiaffonati 2005 Machine ethics and human ethics: A critical view. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*, pp. 103–104

Michael Anderson and Susan Leigh Anderson 2006 Special Issue on Machine Ethics. *IEEE Intelligent Systems*

Michael Anderson and Susan Leigh Anderson 2007 Machine ethics: creating an ethical intelligent agent. *AI Magazine* **28**(4), pp. 15–58

Douglas R. Hofstadter 2007 *I Am a Strange Loop.* New York: Basic Books

George Lakoff and Mark Johnson 1980 *Metaphors we Live By.* Chicago .University Press

Kathryn Blackmond Lasky and Paul E. Lehner 1994 Metareasoning and the problem of small worlds. *IEEE Trans. Sys., Man, and Cybernetics* **24**(11), pp. 1643–1652

Drew McDermott 2001 *Mind and Mechanism.* Cambridge, Mass.: MIT Press

James H. Moor 2006 The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Sys* **21**(4), pp. 18–21

Dana S. Nau 2007 Current trends in automated planning. *AI Magazine* **28**(4), pp. 43–58

Library of Congress Federal Research Division 2007 *Country Profile: United Arab Emirates (UAE).* Available at lcweb2.loc.gov/frd/cs/profiles/UAE.pdf

L. J. Savage 1954 *Foundations of Statistics.* New York: Wiley

Peter Singer 1993 *Practical Ethics.* Cambridge University Press. 2nd ed

Daniel Weld 1999 Recent advances in AI planning. *AI Magazine* **20**(2), pp. 93–123