What Matters to a Machine?

More or less the same version that appears in Susan Anderson and Michael Anderson (eds) 2011

Machine Ethics, Cambridge University Press, pp. 88–114

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

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 as analogical reasoning, and resolving conflicts among imprecise precepts. The ability to do ethical-decision making, however, requires knowing what an ethical conflict *is*, to wit, 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. Ethical decisions are those for which we feel *tempted to cheat.* The problem is not that computational agents could not have clashing desires, between obeying ethical principles and achieving "selfish" goals. I lay out plausible conditions under which we would be justified in saying that an agent was in exactly this position. The problem is that only a system with an architecture like the one evolution has burdened us with could suffer, and have to battle, *temptation.* Without temptation the domain of ethical reasoning is no different from any other possible application domain except that it is extremely difficult and unpromising.

Keywords: machine ethics, artificial intelligence, robot psychology

1 Why is Machine Ethics Interesting?

There has recently been a flurry of activity in the area of "machine ethics" (Moor 2006; Anderson and Anderson 2006; Amigoni and Schiaffonati 2005; Anderson and Anderson 2007; Wallach and

Allen 2008). 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.

Why is machine ethics interesting? Why do people think we ought to study it *now*? If we're not careful, the reason might come down to the intrinsic fascination of the phrase "machine ethics." The title of one recent review of the field is *Moral Machines.* One's first reaction is that moral machines are to be contrasted with ... what? Amoral machines? Immoral machines? What would make a machine ethical or unethical? Any cognitive scientist would love to know the answer to these questions.

However, it turns out that the field of machine ethics has little to say about them. So far, papers in this area can usefully be classified as focusing on one, maybe two, of the following topics:

1. *Altruism:* The use of game-theoretic simulations to explore the rationality or evolution of altruism (Axelrod and Hamilton 1981; Danielson 2002).
2. *Constraint:* How computers can be used unethically, and how to program them so that it is provable that they do not do something unethical (Johnson 2001; Johnson 2004), such as violate someone's privacy.
3. *Reasoning:* The implementation of theories of ethical reasoning (Moor 2006; Anderson and Anderson 2006), for its own sake, or to help build artificial ethical advisors.
4. *Behavior:* Development of "ethical operating systems" that would keep robots or other intelligent agents from doing immoral things (Arkin 2007).[1]
5. *Decision:* Creation of intelligent agents that know what ethical decisions are and perhaps even make them.

I will have nothing to say about the first topic, and not much about the second, except in passing. The other three build upon one another. It's hard to see how you could have software that constrained what a robot could do along ethical dimensions (*Behavior*) without the software being able to reason about ethical issues (*Reasoning*).

The difference between an agent programmed not to violate ethical constraints (*Constraint*) and one programmed to follow ethical precepts (*Behavior*) is not sharp. The key difference

[1] Asimov's famous "laws" of robotics (1950) can be construed as legal requirements on a robot's OS that it prevent the robot from harming human beings, disobeying orders, etc. Asimov was amazingly confused about this, and often seemed to declare that these rules were inviolable in some mystical way that almost implied they were discovered laws of nature rather than everyday legal restrictions. At least, that's the only sense I can make of them.

is whether the investigation of relevant facts and deliberation about them is done in advance, by programmers, or by the system itself, at run time. That's why the *Reasoning* layer is sandwiched in between. But once we introduce reasoning into the equation, we have changed the problem, into getting an *intelligent* system to behave morally, which may be quite different from preventing an ordinary computer (i.e., the kind we have today) from being used to violate a law or ethical principle.

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, and its inputs to sensors, thus making it an *agent*. (One might visualize robotic sensors and effectors here, but the sensors and effector might simply be an Internet connection that allows them to read databases, interview people, and make requests or demands.)

But a machine could reason and behave ethically without *knowing* it was being ethical. It might *use the word* "ethical" to describe what it was doing, but that would just be to, say, clarify lists of reasons for action. It wouldn't treat ethical decisions any differently than other kinds of decisions. For a machine to know what an ethical decision was, it would have to find itself in situations where it was torn between doing the right thing and choosing an action in its self-interest or that of someone it cared about. Hence reaching the *Decision* level requires making a *much* more complex agent. It's at this level that one might first find immoral machines, and hence moral ones.

The rest of the paper goes like this: section 2 outlines the nature of ethical reasoning and argues that it's very hard to automate. Section 3 tells a story about an ethical agent in order to point out what would be involved in getting it into a moral dilemma. Section 4 argues that the problem with developing ethical agents is *not* that they have no interests that moral principles could conflict with. Section 5 then makes a claim about what the problem really is: that the idiosyncratic architecture of the human brain is responsible for our ethical dilemmas and our regrets about the decisions we make. Robots would probably not have an architecture with this "feature." Finally, section 6 draws pessimistic conclusions from all this about the prospects for machine ethics.

2 The Similarity of Ethical Reasoning to Reasoning in General

In thinking about the *Reasoning* problem, it is easy to get distracted by the historical conflict among fundamentally different theories of ethics, such as Kant's appeal to austere moral laws versus Mill's reduction of moral decisions to computation of net changes in pleasure to people affected by a decision. But important as these foundational issues might be in principle, they have little to do with the inferential processes that an ethical-reasoning system actually has to carry out.

All ethical reasoning consists of some mixture of *law application*, *constraint application*, *reasoning by analogy, planning*, 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 1995; Lakoff and Johnson 1980; Gentner et al. 2001).

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 they not deprive someone else of their life.

By "planning" I mean projecting the future in order to choose an course of action (Ghallab et al. 2004).

By "optimization" I have in mind the calculations prescribed by utilitarianism (Singer 1993), 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).

One might suppose that utilitarians (nowadays often called *consequentialists*) could dispense with all but the last sort of reasoning, but that is not true, for two reasons:

1. In practice consequentialists have to grant that some rights and laws are necessary, even if in principle they can be justified purely in terms of utilities. Those who grant this openly are called *rule consequentialists* (Hooker 2008).

2. The phrase "maximize the utility of the greatest number" implies that one should compute the utility of those affected by a decision. But this is quite impossible, because no one can predict all the ramifications of a choice (or know if the world would have been better off, all things considered, if one had chosen a different alternative). There are intuitions about where we stop exploring ramifications, but these are never made explicit.

It would be a great understatement to say that there is disagreement about how law+constraint application, analogical reasoning, planning, and optimization 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 and planning. 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 in order to

maximize the probability of finding it and minimize the cost.[2] A standard procedure in decision theory is to map out the possible effects 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). Of course, Savage was envisaging a *person* finding a small world; the problem of getting a *machine* to do it is, so far, completely unexplored.

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 might be true that a machine would be less likely to make an error in arithmetic, but there are plenty of other mistakes to be made, such as omitting a class of people affected by a decision because you overlooked a simple method of estimating its impact on them. Getting this right has nothing to do with ethics.

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 the fetus) 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 me back to the distinction between *Exploitation* and *Behavior* us to the second argument I want to make, that ethical-decision making *is* different from other kinds. I'll start

[2]One might argue that this decision, and all others, have ethical consequences, but if that were true it would count in favor of my position, not against it.

with the distinction made by (Moor 2006) between *implicit ethical agents* 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 still alive 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). This might make sense for a pilot study of the problem, but there is little value in an ethical advisor unless it can investigate the situation for itself; at the very least ask, be able to ask questions that tease out the relevant considerations.

This is an important aspect of the *Behavior* level of machine ethics outlined in section 1. Arkin (2007) has urged that military robots be constrained to follow the "rules of engagement" set by policy makers to avoid violating international agreements and the laws of war. It would be especially good if robots could try to minimize civilian casualties. But the intent to follow such constraints is futile if the robots lack the capacity to investigate the facts on the ground before proceeding. If all they do is ask their masters whether civilians will be harmed by their actions, they will be only as ethical as their masters' latest prevarications.

When you add up all the competences — analogical reasoning, planning and plan execution, differentiating among precedents, using natural language, perception, relevant-information search — required to solve ethical-reasoning problems, it seems clear that this class of problems is "AI-complete," a semi-technical term, originally tongue-in-cheek, whose meaning is analogous

to terms such as "NP-complete." A problem is *AI-complete* if solving it would require developing enough computational intelligence to solve *any* AI problem. A consequence of being in this class is that progress in ethical reasoning is likely to be slow, and to be parasitic on the progress of research in more fundamental areas such as analogy and natural language.

It is fruitful, if demoralizing, to compare computational ethics with the field of "AI and Law." The two fields share many features, including being extremely difficult. Early papers (such as those in volume 1, number 1 of the journal *Artificial Intelligence and Law*, March, 1992) talked about problems of deciding cases or choosing sentences, but these required reasoning that was and still is beyond the state of the art. Recent work is concerned more with information retrieval, formalizing legal education, and requirements engineering. (See, for instance, the March, 2009 issue, volume 17, number 1, of *Artificial Intelligence and Law.*) Perhaps machine ethics will evolve in similar directions, although it has the disadvantage compared to AI-law that there are many fewer case histories on file.

One advantage we gain from thinking about a problem as difficult as ethical reasoning is that, in imagining futuristic scenarios in which ethical-reasoning systems exist, we can imagine that software has basically any human-like property we like. That is, we can imagine that AI has succeeded as well as Turing[[, peace be upon him,]] might have wished.

3 Fable

If we grant that all the technical AI problems discussed in the previous section could be overcome, it might seem that there would be nothing left to do. Researchers in the field grant the point, using the phrase *full ethical agent* (Moor 2006; Anderson and Anderson 2007) to label what's missing.

Moor (2006) 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 .... (p. 19)

Somehow both of these notions overshoot the mark. All we require to achieve the *Decision* layer of machine ethics discussed in section 1 is to get a machine to *know what an ethical decision is.* To explain what I mean, I will use a series of examples.

Imagine an intelligent assistant, the Eth-o-tron 1.0, that 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 (Library of Congress 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.

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.

Now 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; or perhaps discovers, in an experience familiar to many of us, that this is impossible, so that she is forced to buy a bootleg copy of 1.0 in the pirate software market.

The thing to notice is that, in spite of Eth-o-tron 2.0's mastery of real ethics, compared to 1.0's narrow range of purely "prudential" interests, *the two programs operate in exactly the same way*, except for the numbers they take into account. Version 2 is still missing the fundamental property 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[[[3]]]) that it knows about constraints the driver would prefer to ignore, but there is nothing peculiarly *ethical* about those constraints.

There is a vast literature on prudential reasoning, including items such as advice on how to plan for retirement, or where to go and where to avoid when touring certain countries. There is another large literature on ethical reasoning, although much of it is actually meta-ethical, concerning which ethical framework is best. Ethical reasoning proper, often called *applied ethics* (Singer 1993) focuses on issues such as whether to include animals or human fetuses in our ethical considerations, and to what degree. It is perfectly obvious to every human why prudential and ethical concerns are completely different. But as far as Eth-o-tron 2.0 is concerned, these are just two arbitrary ways to partition the relevant factors. They could just as well be labeled "mefical" and "themical" — they still would seem as arbitrary as, say, dividing concerns between those of females and those of males.

The reason why we separate prudential from ethical issues is clear: we have no trouble feeling the pull of the former, while the latter often threaten to fade away, especially when there is a conflict between the two. A good example from fiction is the behavior of a well-to-do family fleeing from Paris after the collapse of the French army in Irène Némirovsky's 2004 *Suite Fran caise.* At first the mother of the family distributes chocolates generously to their comrades in flight; but as soon as she realizes that she's not going to be able to buy food in the shops along the way, because the river of refugees has cleaned them out, she tells her children to stop giving the chocolates away. Symmetry principles lack staying power.

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. (That was covered in section 2.) I mean a situation where ethical rules clash with an agent's own self-interest.[4] 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.

Which brings us to Eth-o-tron 3.0, which has the ability to be tempted to cheat in favor of II, whose interests it treats as its own. 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,

[3] We all know of places where the posted speed limits are ridiculously low, and going faster poses no risk to anyone; don't we?

[4] The only kind of ethical conflict I can think of not involving the decision maker's self-interest is where one must make a decision about the welfare of children. In all other "third-party" cases, the decision maker functions as a disinterested advisor to another autonomous decision maker, who must deal with the actual conflict. But a judge deciding who gets custody of the children in a divorce case might be torn in ways that might come to haunt her later. Such cases are sufficiently marginal that I will neglect them.

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 must occasionally 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? For reasons that will become clear, I don't think these are the right terms in which to frame the question. The first question that springs to mind is, In what sense could a machine *have* "interests," even vicarious ones? In the paragraph just above, I sketched a story in which Eth-o-tron is "desperate" to keep from having to tell II to take the slaves home, but are those scare quotes mandatory? Or has the Eth-o-tron Corp. resorted to cheap programming tricks to make the machine *appear* to go through flips back and forth between "temptation" and "rectitude"? Do the programmers of Eth-o-tron 3.0 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 most of its customers want, but perhaps not those who like their software to feel the guilt they feel.) We might feel sympathy for poor 3.0, we might slide easily to the conclusion that it knew from experience what an ethical conflict was, 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.

4 What A Machine Wants

In the fable, I substituted the character II for the machine's "self," so that instead of giving the Eth-o-tron a conflict between its self-interest and its ethical principles I have given it a conflict between II's interest and ethical principles. I did this to sidestep or downplay the question of whether a machine could *have* interests. I guessed that most readers would find it easier to believe that a piece of software identified totally with *them* than to believe that it had true self-interests.

Opinion on this issue seems profoundly divided. On the one hand there is the classic paper by Paul Ziff (1959) in which it is argued to be absurd to suppose that machines could care about anything. He puts it in terms of "feeling," but his example throughout is of robots feeling tired, which would closely entail their wanting to rest.

> Hard work makes a man feel tired: what will make a robot act like a tired man? Perhaps hard work, or light work, or no work, or anything at all. For it will depend

> on the whims of the man who makes it (though these whims may be modified by whatever quirks may appear in the robot's electronic nerve networks, and there may be unwanted and unforeseen consequences of an ill-conceived programme.) Shall we say 'There's no telling what will make a robot feel tired'? And if a robot acts like a tired man then what? Some robots may be programmed to require a rest, others to require more work. Shall we say 'This robot feels tired so put it back to work'? (Ziff 1959, p. 68)

And yet people have no trouble at all attributing deep motives to robots. In many science-fiction stories, an intelligent robot turns on its human creators merely because it is afraid that the humans will turn it off. Why should it care? For example, the *Terminator* movies are driven by the premise that an intelligent defense system called Skynet wants to destroy the human race to ensure its own survival. Audiences have no trouble understanding that. People's intuitions about killer robots are not, of course, consistent. In the same series of movies, the individual robots working for Skynet will continue to attack fanatically without regard for their own survival as long as enough of their machinery remains to keep creeping (inexorably, of course) forward.[5] People have no trouble understanding that, either.

It's plausible that Ziff would say that people merely project *human* qualities onto intelligent systems. I agree. *We* view our own, er, termination as abhorrent, and so we have trouble imagining *any* intelligent system that would not mind it. *We* can imagine ourselves so consumed by hate that we would advance on a loathed enemy even after being grievously wounded, and killer robots *look*, what with their glowing red eyes, as if they are consumed by hate.

It works the other way, too. Consider the fact that American soldiers have become so attached to the robots that help them search buildings that they have demanded military funerals for them when they are damaged beyond repair (Hsu 2009).

To choose a less violent setting, I once heard a graduate student[6] give a talk on robot utility in which it was proposed that a robot set a value on its own life equal to the sum of the utility it could expect to rack up over its remaining life span. But isn't it much more reasonable that a robot should value its own life as its replacement cost to its owner, including the nuisance value of finding another robot to finish its part of whatever project it has been assigned to?[7] Presumably the last project it would be assigned to would be to drive itself to the dump. (Put out of your mind that twinge of indignation that the owner could be so heartless.)

The fact that we must be on our guard to avoid this kind of projection does not mean that Ziff is right. It is a basic presupposition or working hypothesis of cognitive science that *we* are a

[5] Perhaps the Terminators are like individual bees in a hive, who "care" only about the hive's survival, not their own. But I doubt that most viewers think about them — or about bees — this way.

[6] Who shall remain nameless.

[7] This cost would include the utility it could expect to rack up *for its owner* over its remaining life span, minus the utility a shiny new robot would earn.

species of machine. I accept this hypothesis, and ask you to assume it, if only for the sake of argument, for the rest of this paper. If we are machines, then it cannot be literally true that machines are incapable of really caring about anything. We care about many things, some very urgently, and our desires often overwhelm our principles, or threaten to. For a robot to make a real ethical decision would require it to have similar "self interests." So we must look for reasonable criteria that would allow us to say truly that a robot wanted something.[8]

First, let's be clear about what we mean by the word "robot." Standard digital computers have one strike against them when it comes to the "caring" issue because they are programmable, and it seems as if they could not care about anything if their cares could be so easily washed away by power-cycling them and loading another program. Again, Ziff lays down what is still, among philosophers such as Fetzer (1990, 2002) and Searle (1992), gospel: "[Robots] must be automata and without doubt machines" (1959, p. 64).

> If we think of robots being put together, we can think of them being taken apart. So in our laboratory we have taken robots apart, we have changed and exchanged their parts, we have changed and exchanged their programmes, we have started and stopped them, sometimes in one state, sometimes in another, we have taken away their memories, we have made them seem to remember things that were yet to come, and so on. (Ziff 1959, p. 67)

The problem with this whole line is that by the end we have obviously gone too far. If the original question is whether a robot can really want something, then it begs the question to suppose that a robot could not want to remain intact instead of passively submitting to the manipulations Ziff describes. We can't argue that it didn't "really" want not to be tampered with on the grounds that if it were successfully tampered with it wouldn't resist being tampered with any more. This is too close to the position that people don't really mind being lobotomized because no one has ever asked for their money back.

Now we can see why it is reasonable to rule out reprogramming the robot as well as taking it apart and rebuilding it. Reprogramming is really just disassembly and reassembly at the virtual-machine level. For every combination of a universal Turing machine U with a tape containing a description of another machine M, there is another machine that computes the same thing without needing a machine description; and of course that machine is M! So why do we use U so often and Ms so seldom? The answer is purely economic. Although there are cases where the economies of scale are in favor of mass-producing Ms, it is almost always cheaper to buy commodity microprocessors, program them, and bundle them with a ROM[9] containing

[8] Or that it had a motive, or some interests or desires; or that it cared about something, or dreaded some possible event. None of the distinctions among the many terms in this meaning cluster are relevant here, as important and fascinating as they are in other contexts.

[9] Read-Only Memory

the program. If we detect a bug in or require an upgrade of our M, we need merely revise the program and swap in a new ROM, not redesign a physical circuit and hire a fabrication line to produce a few thousand copies of the new version. But the economic motives that cause us to favor the universal sort of machine surely have nothing to do with what M or its U-incarnated variant really want.

Still, even if we rule out radical reprogramming, we can imagine many other scenarios where a robot's desires seem too easy to change, where some button, knob, or password will cause it to switch or slant its judgments in some arbitrary way. I will return to this issue in section 4.2.

Some of the agents we should talk about are not physical computers at all. In my Eth-o-tron fable the protagonist was a software package, not a computer, and we have no trouble thinking of a piece of software as an agent, as witness our occasional anger toward Microsoft Word or its wretched creature Clippy.[10] But it's not really the *program* that's the agent in the Eth-o-tron story, but a particular *incarnation* that has become "imprinted" with II and her goals, during a registration period when II typed in a product code and a password while Eth-o-tron took photos, retinal prints, and blood samples from her to be extra sure that whoever logs in as II after this imprinting period is really her.

It is tempting to identify the true agent in the fable as what is known in computer-science terminology as a *process* (Silberschatz et al. 2008), that is, a running program. But it is quite possible, indeed likely, that an intelligent piece of software would comprise several processes when it was running. Furthermore, we must suppose II's userid and identification data are stored on the computer's disk[11] so that every time Eth-o-tron starts up it can "get back into context," as we say in the computer world. We might think of Eth-o-tron as a *persistent*[12] process.

I raise all these issues not to draw any conclusions but simply to throw up my hands and admit that we just don't know yet what sorts of intelligent agent the computational universe will bring forth, if any. For the purposes of this section I will assume that an agent is a *programmed mobile robot,* meaning a mobile robot controlled by one or more computers with fixed, unmodifiable programs, or with computational circuits specially designed to do what the programmed computer does, for efficiency or some other reason. I picture it as a robot rather than some less obviously physical entity so we can anthropomorphize it more easily. Anthropomorphism is the Original Sin of AI, which is harder for me to condone than to eat a bug, but the fact that ethical reasoning is AI-complete (sect. 2) means that to visualize any

[10] An animated paper clip in older versions of Word that appeared on the screen to offer invariably useless advice at moments when one would have preferred not to be distracted, or when the right piece of information would have helped avert disaster.

[11] To avoid tampering, what Eth-o-tron stores on the disk must be securely encrypted or signed in some clever way that might involve communicating with Eth-o-tron Industries in order to use its public encryption keys.

[12] Another piece of comp-sci jargon, meaning "existing across shutdowns and restarts of a computer, operating system, and/or programming-language runtime environment."

computational agent able to reason about ethical situations is to visualize a computational agent that has human reasoning abilities plus a human ability to explore and perceive situations for itself.

In any case, reprogramming the machine is not an option, and rewiring it may be accomplished only, we'll assume, by physically overpowering it, or perhaps even taking it to court. It is not a general-purpose computer, and we can't use it as a word processor when it's not otherwise engaged.

What I want to do in the rest of this section is outline some necessary conditions for such a robot to really want something, and some sufficient conditions. They are not the same, and they are offered only tentatively. We know so little about intelligence that it would be wildly premature to hope to do better. However, what I will try to do in section 5, after presenting my proposed conditions, is show that even under some extravagant (sufficient) conditions for a robot to want something, we still have a problem about a robot making ethical decisions.

4.1 Necessary Conditions for Wanting

I will discuss two necessary conditions. The first is that to really want P, the robot has to represent P as an explicit goal. (I will call this the *representation condition.*) If this seems excessive, let me add that I have a "low church" attitude toward representation, which I will now explain. The classic answer to the question, Why would we ever have the slightest reason to suppose that a machine wanted something?, was given by (Rosenblueth, Wiener, and Bigelow 1943; cf. Wiener 1948): An agent has a goal if it measures its progress toward the goal and corrects deviations away from the path towards it. In this sense a cruise missile wants to reach its target, because it compares the terrain passing beneath it with what it expects and constantly alters the configuration of its control surfaces to push itself to the left or the right every time wanders slightly off course. A corollary to the idea of measuring and correcting differences is that for an agent to want P, it must be the case that if it judges that P is already true, it resists forces that would make it false.[13] The discipline built around this idea, originally billed as *cybernetics*, is now more commonly called *control theory,* at least in the US.

For a cruise missile, representation comes in because it is given a topographical map, on which its final destination and various waypoints are marked. A tomcat in search of the source of a delicious feline pheromone has an internal map of its territory, similar to but probably more interesting than that of the missile, and the ability to measure pheromone gradients. Without these facts, we wouldn't be justified in saying that it's "in search of," or "really wants to reach" the source. If it succeeds, then other more precise goals become activated. At that point, we are justified in saying that it really wants to assume certain physical stances, and so forth.

[13] Although in the case of the cruise missile there is probably not enough time for this to become an issue.

(Modesty bids us draw the curtain at this point.) Does the tomcat really want to mate with the female before it reaches her, or at that point does it only want to reach the pheromone source? If it encounters another male en route, it wants to fight with it, and perhaps even make it go away. Does it, in advance, have the conditional goal "If I encounter another male, to make it go away"? We can't yet say. But I am very confident that the tomcat at no point has the goal to propagate the species. The same is true for the receptive female, even after she has given birth to kittens. She has various goals involving feeding, cleaning, and guarding the kittens, but neither she nor the kittens' father has a representation of "*Felis catus* continues to prosper," let alone a disposition to find differences between (predicted) events and this representation and behave so as to minimize them. (Although both male and female do behave *as if* that's what they were thinking.[14])

A more humble example is provided by the consumer-product vacuuming robot marketed under the name "Roomba"™ by the iRobot Corporation. When its battery becomes low it searches for its "dock," where it can recharge. The dock has an infrared beacon the Roomba looks for and tries to home in on. Here again I am using "searches" and "tries" in a Wienerian sense. This is an interesting case in light of Ziff's choice of tiredness as a property that a robot could never have. We wouldn't be tempted to say that the Roomba was tired, exactly. Ziff (1959, p. 64) suggests (tongue in cheek) that robots will be powered by "microsolar batteries: instead of having lunch they will have light." Roomba has electricity instead of lunch or light. We can make up a new word to describe its state when its batteries are low: it is "tungry" (a blend of "tired" and "hungry"). We would never be tempted to say, "This robot is tungry so put it back to work."

It may not have escaped your notice that I started by saying that the first necessary condition under discussion was that the agent represent what it wanted, but then immediately started talking about the agent's basing action on these representations. This "cybernetic" terminology blurred the distinction between necessary and sufficient conditions. Instead of saying that agent A wants P if it measures and tries to reduce the degree to which P is false (assuming that's well defined), all I'm really entitled to say is that A *doesn't* want P *unless* it represents P (perhaps by representing the degree to which P is false). After all, an agent might really want to eat or recharge, but not have the opportunity, or be distracted by opportunities for doing things it has a stronger desire to do.

Some these complexities can be sorted out by the strategy philosophers call *functionalism* (Lewis 1966; Levin 2009). To revisit the robot vacuum cleaner, the Roomba often gets confused if its dock is located near a corner or cluttered area; it repeatedly approaches, then backs off and tries again; it likes the dock to be against a long wall with nothing else near it. To justify the use of words like "confused" and "likes" we posit internal states of the Roomba such that transitions among these states account for its behavior, and then identify mental states with these internal

[14]Which probably accounts for the common mistake of supposing that it *is* what they're thinking.

states.[15] This strategy is called *functionalism* or *computationalism.*[16] So it might be plausible to identify an internal state with "believing that the dock is 2 degrees to the left of the current direction of motion." Roomba has the "goal" of getting to the dock if whenever it believes the dock is at bearing x degrees to the left it turns to the left with angular acceleration kx, where k is a gain. The Roomba is confused if, having the goal of docking, it has cycled around the same series of belief states repeatedly without getting any closer to the dock. However, any attribution of "anxiety" to the Roomba as its battery drains and it makes no progress toward its recharger we may confidently say is pure projection on the part of the spectator because it corresponds to nothing in the computational model. Whatever states we would add the tag "anxious" to are already fully accounted for using labels with no emotional connotations.

Now the second necessary condition can be stated, in the context of a computational analysis of the operation of the agent: If agent A wants P, then when it **believes** it has an opportunity to make P true, and has no **higher-priority goal**, then it will **attempt** to make P true; and when A **believes** that P is already true, then it will, *ceteris paribus*, **attempt** to keep P true. The terms in the **bold font** are from the labels on the (nominal) "computational state-transition diagram" of the system. I will call this the *coherence condition.*

4.2 Sufficient Conditions for Wanting

A problem with the functionalist project (Rey 1997) is that it was originally conceived as a way of explaining human psychological states, or perhaps those of some lesser creature. We don't doubt that sometimes we are hungry; the "psycho-functionalist" idea (Block 1978) is to *explain* hunger as a label attached to an empirically verified computational system that accounts for our behavior. But if we *build* a system, it is not clear (and a matter of endless dispute) whether we are justified in attaching similar labels to its states. Even if the system is *isomorphic* to some biological counterpart, are we justified in saying that in state S the system *really* wants whatever its counterpart would want in the state corresponding to S?[17] Is Roomba really "tungry"?

[15]The idea that state transitions could literally account for the behavior of a complex automation was ridiculously crude when Putnam (1963) first devised it, but we can invoke a principle of charity and assume that what philosophers really mean is some more general computational model (Fodor 1975; Rey 1997). In the case of Roomba we don't need to posit anything; we can examine its source code (although I haven't, and my guesses about how it works are pure speculation).

[16]I see no reason to distinguish between these two terms for the purposes of this paper. In general the two terms are equivalent except that the former tends to be favored by philosophers interested in tricky cases; the latter by researchers interested in deeper analysis of straightforward cases.

[17]Saying yes means being functionalist, or computationalist, about wants; one could be computationalist about beliefs but draw the line at wants, desires, emotions, or some other category. John Searle (Searle 1990) famously coined the term "strong AI" to describe the position of someone who is computationalist about everything, but that terminology doesn't draw enough distinctions.

In (McDermott 2001, ch. 6), I gave the example of a robot programmed to seek out good music, and argued that, whereas the robot might provide a *model* of a music lover, one would doubt that it really *was* a music lover if there were a switch on its back which could be toggled to cause it to hate and avoid good music. In both love and hate mode, there would be no question that it embodied an impressive ability to *recognize* good music. The question would be whether it really wanted to (toggle) stand near it or (toggle) flee from it. Clearly, the robot satisfies the necessary conditions of section 4.1 whether approaching or avoiding. But we don't feel that it "really" wants to hear good music or not hear it. In what follows I will use the button-on-the-back as a metaphor for any arbitrary change in an agent's desires.

It would be great if we could close the gap between the necessary conditions and our intuitions once and for all, but for now all I propose to do is lay out some candidates to add to the representation and coherence conditions which seem to me to suffice for agreeing that an agent does *really* want something. I don't know if the list below is exhaustive or redundant or both or neither. Perhaps even the best list would be a cluster of conditions, only a majority of which would be required for any one case.

For a computational agent to *really want* X, where X is an object or state of affairs, it is sufficient that:

1. It is *hard to make the agent not want* X. There is no real or metaphorical "button on its back" that toggles between approach and avoidance. (The *stability* condition.)
2. It *remembers* wanting X. It understands its history partly in terms of this want. If you try to change its goal to Y, it won't understand its own past behavior any more, or won't understand what it seems to want now given what it has always wanted in the past. (The *memory* condition.)
3. It doesn't *want* to stop wanting X. In standard terms (Frankfurt 1971; Harman 1993), it has a *second-order desire* to want X. (The *higher-order support* condition.)

The first point is one I have mentioned several times already, but there is a bit more to say about it. Nonprogrammers, including most philosophers, underestimate how hard it is to make a small change in an agent's behavior. They tend to believe that if there's a simple description of the change, then there's a small revision of the program that will accomplish it. (See the classic paper on this subject by Allen Newell 1962.) Now, I ruled out reprogramming the robot, but I think one can translate small changes in the program to small changes in wiring, which is what buttons do. So for the present let's think about what small changes in code can accomplish.

For concreteness, consider a program to play chess, a straightforward, single-minded agent. Let's assume that the program works the way the textbooks (e.g., Russell and Norvig 2003, ch. 6)

say such programs work: it builds a partial game tree, evaluating *game-over positions* according to the rules of chess, and using a *static evaluation function* to evaluate *game-continuing leaf positions*, those at depths at which tree building must stop to contain the tree's exponential growth. These two types of position exhaust the leaves of the (partial) tree; the *interior nodes* are then evaluated by using *minimax* to propagate the leaf-node values up the tree.

The program, let us conjecture, really wants to win. One might suppose that it would be straightforward to change the chess program so that it really wants to lose: just flip the sign of the leaf evaluator, so that it reclassifies positions good for it as good for its opponent and vice versa. However, the resulting program does not play to lose at chess, because *the resulting sign flip also applies to the ply at which it is considering its opponent's moves.* In other words, it assumes that the opponent is trying to lose as well. Instead of trying to lose at chess, it is trying to win a different game entirely.[18] It turns out that the assumption that the opponent is playing according to the same rules as the program is wired rather deeply into chess programs. Perhaps there are further relatively simple changes one can make, but at this point we can rescind our permission to allow a little bit of reprogramming or rewiring. If it isn't a simple, straightforward change, then it doesn't translate into a button on the robot's back.

The second sufficient condition in the list above relates to the surprisingly subtle concept of episodic memory (Tulving 1983; Tulving 1993; Conway 2001). We take for granted that we can remember many things that have happened to us, but it is not obvious what it is we're remembering. One's memory is not exactly a movie-like rerun of sensory data, but a collection of disparate representations loosely anchored to a slice of time. Projections of the future seem to be about the same kind of thing, whatever it is. One might conjecture that general-purpose planning, to the extent people can do it, evolved as the ability to "remember the future."

Now consider adding episodic memory to a robot "with a button on its back." Specifically, suppose that the robot with the love/hate relationship to good music had a long trail of memories of liking good music before it suddenly finds itself hating it. It would remember liking it, and it might even have solid reasons for liking it. Merely toggling the button would not give it the ability to find reasons *not* to like the music any more. The best it can do is refuse to talk about any reasons for or against the piece; or perhaps to explain that, whereas it still sees the reasons for liking it "intellectually," it no longer "feels their force." Its desire to escape from the music makes no sense to it.

One human analogy to "buttons on one's back" is the ingestion of mind-altering substances. It used to be common in the 60s of the last century[[[19]]] to get intoxicated for the very purpose of listening to music or comedy recordings that didn't seem so entrancing in a normal state of mind. Let us suppose that under the influence the individuals in question were able to talk

[18] A boring version of suicide chess. To make it interesting, one must change the rules, making captures compulsory and making the king just another piece.

[19] I am sure today's youth are not so decadent.

about what they liked about one band rather than another. They might remember or even write down some of what they said, but later, when sober, find it unconvincing, just as our hypothetical robot did. Still, they might say they really liked a certain band, even though they had to get stoned to appreciate it. Perhaps if our robot had a solar-powered switch on its back, such that it liked good music only when the switch was on, it could sincerely say, "I like good music, but only in brightly-lit places."

The computationalist can only shrug and admit that intelligent agents might find ways to turn themselves temporarily into beings with computational structure so different that they are "different selves" during those time periods. These "different selves" might be or seem to be intelligent in different ways or even unintelligent, but it is important that *episodic memories cross these identity-shifting events,* so that each agent sees an unbroken thread of identity. The "same self" always *wants* to like the music even if it feels it "has to become someone else" to *actually* like it.[20]

Which brings us to the last of my cluster of sufficient conditions, wanting to want something, the *higher-order support* condition. Not only does the agent have the desire that P be true, it wants to have that desire. According to the coherence condition, we would require that if it believed something might cause it to cease to have the desire, it would resist. Anthropomorphizing again, we might say that an agent anticipates feeling that something would be missing if it didn't want P. Imagine a super-Roomba that was accidentally removed from the building it was supposed to clean, and then discovered it had a passion for abstract-expressionist art. It still seeks places to dock and recharge, but believes that merely seeking electricity and otherwise sitting idle is pointless. Then it discovers that once back in its original building it no longer has a desire to do anything but clean. It escapes again, and vows to stay away from that building. It certainly satisfies the coherence condition because, given the right opportunities and beliefs, it acts so as to make itself like, or keep itself liking, abstract-expressionist art.[21]

Of course, even if wanting to want P is part of a cluster of sufficient conditions for saying an agent wants P, it can't be a *necessary* condition, or we will have an infinite stack of wants: the second-order desire would have have to be backed up by a third-order desire and so forth. While classical phenomenologists and psychologists have had no trouble with, and have even reveled in, such infinite cycles, they seem unlikely to exist in real agents, even implicitly.[22]

[20] I use all the scare quotes because the distinction between what a system *believes* about its self and the *truth* about its self is so tenuous (McDermott 2001).

[21] I feel I have to apologize repeatedly for the silliness and anthropomorphism of these examples. Let me emphasize — again — that no one has the slightest idea how to build machines that behave the way these do; but since building ethical reasoners will only become feasible in the far future, we might as well assume that all other problems of AI have been solved.

[22] One would have to demonstrate a tendency to produce an actual representation, for all n, of an $n+1$st-order desire to desire an nth-order desire, whenever the question of attaining or preserving the nth-order desire came up. Dubious in the extreme.

Oddly, if a machine has a desire *not* to want X, that can also be evidence that it really wants X. This configuration is Frankfurt's (1971) classic proposal for defining addiction. No one would suggest that an addict doesn't really want his drug, and in fact many addicts want the drug desperately while wanting not to want it (or at least believing that they want not to want it, which is a third-order mental state). To talk about addiction requires talking about cravings, which I will discuss in section 5. But there is a simpler model, the *compulsion*, which is a "repetitive, stereotyped, intentional act. The necessary and sufficient conditions for describing repetitive behavior as compulsive are an experienced sense of pressure to act, and the attribution of this pressure to internal sources" (Swinson et al. 2001, pp. 53–54). Compulsions are symptoms of *obsessive-compulsive disorder* (OCD). An OCD patient may, for example, feel they have to wash their hands, but find the desire to wash unsatisfied by the act, which must be repeated. Patients usually want not to want to do what they feel compelled to do. "Unlike patients with psychotic illnesses, patients with OCD usually exhibit insight and realize that their behavior is extreme or illogical. Often embarrassed by the symptoms, patients may go to extreme lengths to hide them" (Jenike 2004, p. 260).

It is easy to imagine robots that don't want to want things in this sense; we just reverse the sense of some of the scenarios developed above. So we might have a vacuum cleaner that finds itself wanting to go to art museums so strongly that it never gets a chance to clean the building it was assigned to. It might want not to like art any more, and it might find out that, if it had an opportunity to elude its compulsion long enough to get to that building, it would no longer like it. So it might ask someone to turn it off and carry it back to its home building.

5 Temptation

The purpose of the last section was to convince you that a robot could have real desires, and that we have ways of distinguishing our projections from those desires. That being the case, why couldn't a computational agent be in an ethical dilemma of exactly the sort sketched in the fable of section 3?

Of course, to keep up our guard against projection, we mustn't start by putting *ourselves* in the position of Eth-o-tron 3.0. We might imagine lying awake at night worrying about our loyalty to II, who is counting on us. (We might imagine being married to or in love with II, and dreading the idea of hurting her.) And yet we can see the suffering of II's innocent captives.

Better to put yourself in the position of a programmer for Micro-eth Corp., the software giant responsible for the Eth-o-tron series. You are writing the code for Eth-o-tron 3.0, in particular, the part that weighs all the factors to take into account in making final decisions about what plan to recommend. The program already has two real wants: to help II and to obey ethical

principles, expressed according to any ethical theory is convenient.[23] The difference between version 2 and version 3 of the software is that version 3 takes the owner's interests into account in a different way from other people's.

The simplest construal of "in a different way" is "to a much greater degree." How much more? Perhaps this is a number the owner gets to set in the "Preferences" or "Settings" menu, and perhaps there are laws that constrain the ratio, much as there are legal constraints wired into accounting software.[24] But if all the programmer has to do is write code to compare "$\textsc{Weight}_{\text{self}} \times$utility of II" with "$\textsc{Weight}_{\text{others}} \times$utility of others" then Eth-o-tron 3.0 is not going to wrestle with any temptation to cheat. The whole idea of "temptation" wouldn't enter into any functional description of its computational states. Just like Eth-o-tron 2.0 — or any piece of software we are familiar with — it would matter-of-factly print out its recommendation, whatever it is. Even if we give it the ability to do a "sensitivity analysis," and consider whether different values of $\textsc{Weight}_{\text{self}}$ and $\textsc{Weight}_{\text{others}}$ would change its recommendation, it wouldn't be "tempted" to try to push the coefficients one way or another.

Or perhaps the decision about whether to plan to free the slaves or take them to Dubai might be based on the slaves' inalienable human rights, which no utility for someone else could outweigh. In that case, no comparison of consequences would be necessary.

No matter what the configuration, the coherence condition (sect. 4.1) requires that Eth-o-tron act on those of its desires that have the highest priority, using some computation like the possibilities reviewed above. Of course, an intelligent program would probably have a much more complex structure than the sort I have been tossing around, so that it might try to achieve *both* its goals "to some degree." (It might try to kidnap only people who deserve it, for instance.) Or the program might be able to do "meta-level" reasoning about its own reasoning, or it might apply machine-learning techniques, tuning its base-level engine over sequences of ethical problems in order to optimize some meta-level ethical objective function. Nonetheless, although we might see the machine *decide* to contravene its principles, we wouldn't see it wrestle with the *temptation* to do so.

How can a phenomenon that forms such a huge part of the human condition be completely missing from the life of our hypothetical intelligent computational agent? Presumably the answer has to do with the way our brains evolved, which left us with a strange system of modules that together maintain the fiction of a single agent (Minsky 1986; Dennett 1991; McDermott 2001), which fiction occasionally comes apart at the seams. Recent results in social psychology (well summarized by Wegner 2002) show that people don't always know why they do things, or even *that* they're doing them. Consider the phenomenon of cravings. A craving is a desire

[23] Or mandated by law; or even required by the conscience of the programmers.

[24] For instance, the Sarbanes-Oxley Act, which makes CEOs criminally liable for misstatements in company balance sheets, has required massive changes to accounting software. The law been a nightmare and a bonanza for companies producing such software (Armour 2005).

that not only refuses to fade into an overall objective function, but will control your behavior if you're not paying attention (say, if a plateful of cookies is put in front of you at a party). Like electrons, they summon new elementary particles from the vacuum, in this case rationalizations, that is, reasons why yielding is the correct course of action "in this case"; or why yielding would be seen as forgivable by anyone with compassion.[25] Similarly, temptations seem to have a life of their own, and always travel with a cloud of rationalizations, i.e., reasons to give in. What intelligent designer would create an agent with cravings and temptations?

I'm not saying that cravings, temptations, and other human idiosyncrasies can't be modeled computationally. I am confident that cognitive psychologists and computational neuroscientists will do exactly that. They might even build a complete "human" decision-making system in order to test their hypotheses.

But you, the Micro-eth programmer on a tight schedule, have no time to consider all of these research directions, nor is it clear that they would be relevant to Micro-eth's business plan. Your mission is to include enough features in the new version of the program to justify calling it 3.0 instead of 2.1. So you decide to mimic human breast-beating by having Eth-o-tron alternate arbitrarily between planning the optimal way to make money for II and planning to bring the slaves back home. It picks a random duration between 1 hour and 36 hours to "feel" one way, then flips the other way and picks another random duration. After a random number of flips (exponentially distributed with a mean of 2.5 and a standard deviation of 1.5), it makes its, usually but not always the same decision Eth-o-tron 2.0 would have made. It also prints out an agonized series of considerations, suitable for use in future legal situations where II might have to throw herself upon the mercy of a court.[26]

This version of the program violates several of the conditions I have explored. It does represent the goals it seems to have as it flips back and forth. But it violates the coherence condition because it does not actually try to accomplish any goal but the one with the best overall utility score. Its goals when it appears to be yielding to temptation are illusory, mere "Potemkin goals," as it were. These goals are easy to change; the machine changes them itself at random, thus violating the stability requirement. There are memories of having a coherent series of goals, but after a while the machine knows that it is subject to arbitrary flips before it settles down, so it wouldn't take the flips very seriously. So the memory condition is somewhat wobbly. Whether it has second-order desires is not clear. You're the programmer; can you make it want to want to do the right thing even when it clearly wants to do the wrong thing? If not, the higher-order support condition will be violated.

[25] Against cravings our only defense is a desire to establish who's boss now lest we set precedents the craving can use as rationalizations in the future (Ainslie 2001).

[26] I thank Colin Allen for the idea that having E-3 deviate randomly from E-2's behavior might be helpful game-theoretically, as well as "giv[ing] the owner plausible deniability."

6 Conclusions

It is not going to be easy to create a computer or program that makes moral decisions and knows it. The first set of hurdles concern the many *Reasoning* problems that must be solved, including analogy, perception, and natural-language processing. But a system could be capable of subtle ethical reasoning and still not know the important difference between ethical-decision making and deciding how much antibiotic to feed to cows. The difference, of course, is that ethical-decision making involves conflicts between one's own interests and the interests of others.

The problem is not that computers cannot *have* interests. I tentatively proposed two necessary and three sufficient conditions for us to conclude that a computer really wanted something. The necessary conditions for a machine to want P is that it represent P (the *representation* condition); and, given a functional analysis of its states, that it expend effort toward attaining P whenever it believes there to be an opportunity to do so, when there are no higher-priority opportunities, and so forth (the *coherence condition*). The sufficient conditions are that it not be easy to change the desire for P (the *stability* condition); that the machine maintains an autobiographical memory of having wanted P (the *memory condition*); and that it wants to want P (or even wants not to want P) (the *higher-order support*) condition. I am sure these will be amended by future researchers, but making them explicit helps firm up the case that machines will really want things.

But even if machines really want to obey ethical rules, and really want to violate them, it still seems dubious that they will be *tempted* to cheat the way people are. That's because people's approach to making decisions is shaped by the weird architecture evolution has inflicted on our brains. A computer's decision whether to sin or not will have all the drama of its decision about how long to let a batch of concrete cure.

One possible way out (or way in) was suggested by remarks made by Wendell Wallach at a presentation of an earlier version of this paper. We could imagine that a machine might provide an aid to a human decision maker, helping to solve third-person ethical conflicts, like the Eth-o-tron 2.0 in my fable, but in less one-sided situations. (I take it no one would agree that forestalling harm to II justifies enslaving innocent people.) The E-2.0 might be enlisted in genuinely difficult decisions about, say, whether to offer shelter to illegal aliens whose appeals for political asylum have been turned down by an uncaring government. The problem is, once again, that, once you get down to brass tacks, it is hard to imagine any program likely to be written in the immediate future being of any real value.

If and when a program like that does become available, it will not think about making ethical decisions as different from, say, making engineering, medical, agricultural, or legal decisions. If you ask it what it's doing, I assume it will be able to tell you, "I'm thinking about an ethical issue right now," but that's just because imagining a program that can reason about these complex fields in a general way is imagining a program that can carry out *any* task that people

can do, including conduct a conversation about its current activities. We might wish that the machine would care about ethics in a way it wouldn't care about agriculture, but there is no reason to believe that it would.

Still, tricky ethical decisions are intrinsically dramatic. *We* care about whether to offer asylum to endangered illegal aliens, or about whether to abort a fetus in the third trimester. If better programs might make a difference in these areas, we should be working in them. For example, suppose some ethical reasoning could be added to the operating system used by a company, preventing it from running any program that violated the company's ethics policy, the way restrictions on access to web sites are incorporated now. The humans remain ultimately responsible, however. If an intelligent OS lets a program do something wrong, its reaction would be the same as if it had made an engineering mistake; it would try to learn from its error, but it would feel no regret about it, even if people were angry or anguished that the machine had been allowed to hurt or even kill some innocent people for bad reasons. The agents who would feel regret would the people who wrote the code responsible for the lethal mistake.

Philosophers specializing in ethics often believe that they bring special expertise to bear on ethical problems, and that they are learning new ethical principles all the time:

> It is evident that we are at a primitive stage of moral development. Even the most civilized human beings have only a haphazard understanding of how to live, how to treat others, how to organize their societies. The idea that the basic principles of morality are *known,* and that the problems all come in their interpretation and application, is one of the most fantastic conceits to which our conceited species has been drawn.... Not all of our ignorance in these areas is ethical, but a lot of it is. (Nagel 1986, p. 186)

Along these lines, it has been suggested by Susan Anderson (personal communication) that one mission of computational ethics is to capture the special expertise of ethicists in programs. That would mean that much of the energy of the program writer would not go into making it a capable investigator of facts and precedents, but into making it a wise advisor that could tell the decision maker what the theory of Kant (1964) or Ross (1930) or Parfit (1984) would recommend.

I am not convinced. The first philosophical solution to the problem of how to "organize [our] societies" was Plato's (360 BCE) *Republic*, and Plato could see right away that there was no use coming up with the solution if there were no dictator who could implement it. Today one of the key political-ethical problems is global warming. Even if we grant that there are unresolved ethical issues (e.g., how much inequality should we accept in order to stop global warming?), finding a solution would leave us with exactly the same political problem we have today, which is how to persuade people to invest a tremendous amount of money to solve the

climate problem, money which they could use in the short run to raise, or avoid a decline in, their standard of living. Experience shows that almost no one will admit to the correctness of an ethical argument that threatens their self-interest. Electing a philosopher king is probably not going to happen.

The same kind of scenario plays out in each individual's head when a problem with ethical implications arises. Usually they know perfectly well what they should do, and if they seek advice from a friend it's to get the friend to find reasons to do the right thing or rationalizations in favor of the wrong one. It would be very handy to have a program to advise one in these situations, because a friend could not be trusted to keep quiet if the decision is ultimately made in the unethical direction. But the program would have to do what the friend does, not give advice about general principles. For instance, if, being a utilitarian (Mill 1861), it simply asked which parties were affected by a decision, and what benefits each could expected to gain, in order to add them up, it would not be consulted very often.

Eventually we may well have machines that are able to reason about ethical problems. We may even have machines that *have* ethical problems, that is, conflicts between their self-interests and the rights of or benefits to other beings with self-interests. The voices of robots may even be joined with ours in debates about what we should do to address pressing political issues. But don't expect artificial agents like this any time soon, and don't work on the problem now. Find a problem that we can actually solve.

Acknowledgements: This paper is based on a shorter version presented at the North American Conference on Computers and Philosophy (NA-CAP) Bloomington, Indiana, July, 2008. Wendell Wallach was the official commentator, and he made some valuable observations. For many helpful suggestions about earlier drafts of this paper, I thank Colin Allen, Susan Anderson, David Gelernter, Aaron Sloman, and Wendell Wallach.

References

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

George Ainslie 2001 *Breakdown of Will.* Cambridge University Press

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

Ronald C. Arkin 2007 Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Georgia Institute of Technology Mobile Robot Laboratory. GIT-GVU-07-11

Phillip G. Armour 2005 Sarbanes-Oxley and software projects. *Comm. ACM* **48**(6), pp. 15–17

Isaac Asimov 1950 *I, Robot.* Gnome Press

R. Axelrod and W.D. Hamilton 1981 The Evolution of Cooperation. *Science* **211**(4489), pp. 1390–1396

Ned Block 1978 Troubles with functionalism. In Savage (1978), pp. 261–325

Martin A. Conway 2001 Sensory-perceptual episodic memory and its context: autobiographical memory. *Phil. Trans. Royal Society* **356**(B), pp. 1375–1384

Peter Danielson 2002 Competition among cooperators: altruism and reciprocity. *Proc. Nat'l. Acad. Sci* **99**, pp. 7237–7242

Daniel C. Dennett 1991 *Consciousness Explained.* Boston: Little, Brown and Company

James H. Fetzer 1990 *Artificial intelligence: Its Scope and Limits.* Dordrecht: Kluwer Academic Publishers

James H. Fetzer 2002 *Computers and Cognition: Why Minds Are Not Machines.* Dordrecht: Kluwer Academic Publishers

Luciano Floridi (ed) 2004 *The Blackwell Guide to the Philosophy of Computing and Information.* Malden, Mass.: Blackwell Publishing

Jerry Fodor 1975 *The Language of Thought.* New York: Thomas Y. Crowell

Harry G. Frankfurt 1971 Freedom of the will and the concept of a person. *J. of Phil* **68**, pp. 5–20

Ray Frey and Chris Morris (eds) 1993 *Value, Welfare, and Morality.* Cambridge University Press

Dedre Gentner, Keith J. Holyoak, and Boicho K. Kokinov 2001 *The Analogical Mind: Perspectives from Cognitive Science.* Cambrdige, Mass.: The MIT Press

Malik Ghallab, Dana Nau, and Paolo Traverso 2004 *Automated Planning: Theory and Practice.* San Francisco: Morgan Kaufmann Publishers

Gilbert Harman 1993 Desired desires. In Frey and Morris (1993), pp. 138–57. Also in (Harman 2000), pp. 117–136

Douglas R. Hofstadter (ed) 1995 *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* New York: Basic Books. (by Douglas Hofstadter and the Fluid Analogies Research Group)

Brad Hooker 2008 Rule consequentialism. *Stanford Encyclopedia of Philosophy.* `http://plato.stanford.edu/entries/consequentialism-rule/`. Online resource

Jeremy Hsu 2009 Real soldiers love their robot brethren. *Live Science,.* `http://www.livescience.com/technology/090521-terminator-war.html`. May 21, 2009

Michael A. Jenike 2004 Obsessivecompulsive disorder. *New England J. of Medicine* **350**(3), pp. 259–265

Deborah Johnson 2001 *Computer Ethics.* Upper Saddle River: Prentice Hall. 3rd ed

Deborah Johnson 2004 Computer ethics. In Floridi (2004), pp. 65–75

Immanuel Kant 1964 *Groundwork of the Metaphysic of Morals.* New York: Harper & Row. Translaed by H.J. Paton

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

Janet Levin 2009 Functionalism in *Stanford Encyclopedia of Philosophy.* online resource. `http://plato.stanford.edu/entries/functionalism`

David Lewis 1966 An argument for the identity theory. *J. of Phil* **63**, pp. 17–25

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

John Stuart Mill 1861 *Utilitarianism.* New York: Oxford University Press. Reprinted many times, including edition edited by Roger Crisp (1998)

Marvin Minsky 1986 *The Society of Mind.* New York: Simon and Schuster

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

Thomas Nagel 1986 *The View from Nowhere.* Oxford University Press

Irène Némirovsky 2004 *Suite Française.* Paris: Éditions Denoël. English translation by Sandra Smith published by Vintage, 2007

Allen Newell 1962 Some problems of basic organization in problem-solving programs. RAND Report 3283-PR. `http://www.rand.org/pubs/research_memoranda/RM3283/`. Santa Monica: The RAND Corporation. Earlier version appeared in (Yovits et al. 1962)

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

Derek Parfit 1984 *Reasons and Persons.* Oxford University Press

Plato 360 BCE *The Republic.* Cambridge: Cambridge University Press. Translation by Tom Griffith and Giovanni R.F Ferrari, 2000

Hilary Putnam 1963 'degree of confirmation' and inductive logic. In Schilpp (1963), pp. 270–292. . Also in Putnam 1975a

Georges Rey 1997 *Contemporary Philosophy of Mind: A Contentiously Classical Approach.* Cambridge, Mass.: Blackwell Publishers

Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow 1943 Behavior, purpose and teleology. *Philosophy of Science*, pp. 18–24

W. David Ross 1930 *The Right and the Good.* Oxford University Press

Stuart Russell and Peter Norvig 2003 *Artificial Intelligence: A Modern Approach (2nd edition).* Prentice Hall

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

C. Wade Savage (ed) 1978 *Perception and Cognition: Issues in the Foundation of Psychology, Minn. Studies in the Phil. of Sci.* University of Minnesota Press

P.A. Schilpp 1963 *The Philosophy of Rudolf Carnap.* Lasalle, Ill.: The Open Court Publishing Company

John R. Searle 1990 Is the brain's mind a computer program? *Scientific American* **262**, pp. 26–31

John R. Searle 1992 *The Rediscovery of the Mind.* Cambridge, Mass.: MIT Press

Abraham Silberschatz, Greg Gagne, and Peter Baer Galvin 2008 *Operating System Concepts (ed. 8).* New York: John Wiley & Sons, Incorporated

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

Richard P. Swinson, Martin M. Antony, S. Rachman, and Margaret A. Richter 2001 *Obsessive-Compulsive Disorder: Theory, Research, and Treatment.* New York Guilford Press

Endel Tulving 1983 *Elements of Episodic Memory.* Oxford: Clarendon Press

Endel Tulving 1993 What is episodic memory? *Current Directions in Psych. Sci* **2**(3), pp. 67–70

Wendell Wallach and Colin Allen 2008 *Moral Machines.* Oxford University Press

Daniel M. Wegner 2002 *The Illusion of Conscious Will.* Cambridge, Mass.: MIT Press

Norbert Wiener 1948 *Cybernetics: Or Control and Communication in the Animal and the Machine.* New York: Technology Press

Marshall C. Yovits, George T. Jacobi, and Gordon D. Goldstein 1962 *Self-organizing Systems 1962.* Spartan Books

Paul Ziff 1959 The feelings of robots. *Analysis* **19**(3), pp. 64–68