Ethics for Machines

"A robot may not injure a human being, or through inaction, allow a human to come to harm."

-- Isaac Asimov's First Law of Robotics

The first book report I ever gave, to Mrs. Slatin's first grade class in Lake, Mississippi in 1961, was on a slim volume entitled "You Will Go to the Moon". I have spent the intervening years thinking about the future.

The four decades that have passed have witnessed advances in science and physical technology that would be incredible to a child of any other era. I did see my countryman Neil Armstrong step out onto the moon. The processing power of the computers that controlled the early launches can be had today in a $5 calculator. The genetic code has been broken and the messages are being read -- and in some cases, rewritten. Jet travel, then a perquisite of the rich, is available to all.

That young boy that I was spent time on other things besides science fiction. My father was a minister, and we talked (or in many cases, I was lectured and questioned!) about good and evil, right and wrong, what were our duties to others and to ourselves.

In the same four decades, progress in the realm of ethics has been modest. Almost all of it has been in the expansion of inclusiveness, broadening the definition of who deserves the same consideration you always gave your neighbor. I experienced some of this first hand as a schoolchild in '60's Mississippi. Perhaps the rejection of wars of adventure can also be counted. And yet those valuable advances to the contrary notwithstanding, ethics, and its blurry reflection politics, has seemed to stand still compared to the advances of physical science. This is particularly true if we take the twentieth century as a whole -- it stands alone in history as the "Genocide Century", the only time in history where governments killed their own people by the millions, not just once or in one place but repeatedly, all across the globe.

We can extend our vision with telescopes and microscopes, peering into the heart of the atom and seeing back to the very creation of the universe. When I was a boy, and vitally interested in dinosaurs, no one knew why they had died out. Now we do. We can map out the crater of the Chixulub meteor with sensitive gravitometers, charting the enormous structure below the ocean floor.

Up to now, we haven't had, or really needed, similar advances in "ethical instrumentation". The terms of the subject haven't changed. Morality rests on human shoulders, and if machines changed the ease with which things were done, they did not change responsibility for doing them. People have always been the only "moral agents".

Similarly, people are largely the objects of responsibility. There is a developing debate over our responsibilities to other living creatures, or species of them, which is unresolved in detail, and which will bear further discussion below. We have never, however, considered ourselves to have *moral* duties to our machines, or them to us.

All that is about to change.

What Are Machines, Anyway?

We have a naive notion of a machine as a box with motors, gears, and whatnot in it. The most important machine of the industrial revolution was the steam engine, providing power to factories, locomotives, and ships. If we retain this notion, however, we will fall far short of an intuition capable of dealing with the machines of the future.

The most important machine of the twentieth century wasn't a physical thing at all. It was the Turing Machine, and it was a mathematical idea. It provided the theoretical basis for computers. Furthermore, it established the principle that for higher functions such as computation, it didn't matter what the physical realization was (within certain bounds) -- any computer could do what any other computer could, given enough memory and time.

This theoretical concept of a machine as a pattern of operations which could be implemented in a number of ways is called a virtual machine. In modern computer technology, virtual machines abound. Successive versions of processor chips re-implement the virtual machines of their predecessors, so that the old software will still run. Operating systems (e.g. Windows) offer virtual machines to applications programs. Web browsers offer several virtual machines (notably Java) to the writers of Web pages.

More importantly, any program running on a computer is a virtual machine. Usage in this sense is a slight extension of that in computer science, where the "machine" in "virtual machine" refers to a computer, specifically an instruction set processor. Strictly speaking, computer scientists should refer to "virtual processors", but they tend to refer to processors as "machines" anyway. For the purposes of our discussion here, we can call any program a virtual machine. In fact, I will drop the "virtual" and call programs simply "machines". The essence of a machine, for our purposes, is its behavior; what it does given what it senses (always assuming that there is a physical realization capable of actually doing the actions).

To understand just how complex the issue really is, let's consider a huge, complex, immensely powerful machine we've already built. The machine is the U.S. Government and legal system. It is a lot more like a giant computer program than people realize. Really complex

computer programs are not sequences of instructions; they are sets of rules. This is explicit in the case of "expert systems" and implicit in the case of distributed, object-oriented, interrupt-driven, networked software systems. More to the point, sets of rules are programs -- in our terms, machines.

Of course you will say that the government isn't *just* a program; it's under human control, isn't it, and it's composed of people to begin with. It is composed of people, but the whole point of the rules is to make these people do different things, or do things differently, than they would have otherwise. Indeed in many cases a person's whole function in the bureaucracy is to be a sensor or effector; once the sensor-person does their function of recognizing a situation in the "if" part of a rule (what lawyers call "the facts"), the system, not the person, decides what to do about it ("the law"). Bureaucracies famously exhibit the same lack of common sense as do computer programs.

From a moral standpoint, it is important to note that those governments in the twentieth century which were most evil, murdering millions of people, were autocracies under the control of individual humans such as Hitler, Stalin, and Mao; and that governments which were more autonomous machines, such as the liberal Western democracies, were significantly less evil.

Up to now, the application of ethics to machines, including programs, has been that the actions of the machine were the responsibility of the designer and/or operator. In future, however, it seems clear that we are going to have machines, like the government, whose behavior is an emergent and to some extent unforeseeable result of design and operation decisions made by many people and ultimately by other machines.

Why Machines Need Ethics

Moore's Law is a rule of thumb regarding computer technology which, in one general formulation, states that the processing power per price of computers will increase by a factor of 1.5 every year. This rule of thumb has held true from 1950 through 2000. The factor of a billion improvement in bang-for-a-buck of computers over the period is nearly unprecedented in technology.

Among its other effects, this explosion of processing power, along with the internet, has made the computer a tool for science of a kind never seen before. It is, in a sense, a powered imagination. Science as we know it was based on the previous technology revolution in information, the printing press. The spread of knowledge it enabled, together with the precise imagining ability given by the calculus, gave us the scientific revolution in the seventeenth and eighteenth centuries. That in turn gave us the industrial revolution in the nineteenth and twentieth.

The computer and internet are the calculus and printing press of our day. Our new scientific revolution is going on even as we speak. The industrial revolution to follow hasn't happened yet, but by all accounts it is coming, and well within the twenty-first century, such is the accelerated pace modern technology makes possible.

The new industrial revolution of physical production is sometimes referred to as nanotechnology. On our computers, we can already simulate the tiny machines we will build. They have some of the "magic" of life, which is after all based on molecular machines itself. They will, if desired, be able to produce more of themselves. They will produce stronger materials, more reliable, longer-lasting machines, more powerful motors which are utterly silent, and last but not least, much more powerful computers.

None of this should come as a surprise. If you extend the trend lines for Moore's Law, in a few decades part sizes are expected to be molecular and the price-performance ratios imply something like the molecular manufacturing schemes that nanotechnologists have proposed. If you project the trend line for power-to weight ratio of engines, which has held steady since 1850 going through several different technologies from steam to jet engines, it says we have molecular power plants in the 2030-2050 timeframe.

The result of this is essentially a reprise of the original industrial revolution, a great flowering of increased productivity and capabilities, and a concomitant decrease in costs. In general, we can expect the costs of "hi-tech" manufactured items to follow a downward track as computers have. One interesting corollary is that we will have affordable robots.

Robots today are much more prevalent than people may realize. Your car and your computer were likely partially made by robots. Industrial robots are hugely expensive machines that must operate in a carefully planned and controlled environment, because they have very limited senses and no common sense whatsoever.

With nanotechnology, that changes drastically. Indeed, it's already starting to change, as the precursor technologies such as micromachines begin to have their effect.

Existing robots are often stupider than insects. As computers increase in power, however, they will get smarter, more able to operate in unstructured environments, and ultimately be able to do anything a human can. Robots will find increasing use, as costs come down, in production, in service industries, and as domestic servants.

Meanwhile, because non-mobile computers are already more plentiful and will be cheaper than robots for the same processing power, stationary computers as smart as humans will probably arrive a bit sooner than human-level robots. [see Kurzweil, Moravec]

Before we proceed let's briefly touch on what philosophers sometimes call the problem of other minds. I know I'm conscious, but how do I know that you are -- you might just be like an unfeeling machine, a zombie, producing those reactions by mechanical means. After all, there have been some cultures where the standard belief among men was that women were not conscious (and probably vice versa!). If we're not sure about other people, how can we say that an intelligent computer would be conscious?

This is important to our discussion because there is a tendency for prople to set a dividing line for ethics between the conscious and the non-conscious. This can be seen in formal philosophical treatment as far back as Adam Smith's theory of ethics as based in sympathy. If we can't imagine something as being able to feel a hurt, we have less compunctions about hurting it, for example.

The short answer is that it doesn't matter. [see Dennet, "Intentional Stance"] The clear trend in ethics is for a growing inclusivity in those things considered to have rights -- races of people, animals, ecosystems. There is no hint, for example, that plants are conscious, either individually or as species, but that does not, in and of itself, preclude a possible moral duty to them, at least species of them.

A possibly longer answer is that the intuitions of some people (Berkeley philosopher John Searle, for example) that machines cannot "really" be conscious are not based on any real experience with intelligent machines, and that the vast majority of people interacting with a machine that could, say, pass the unrestricted Turing Test, would be perfectly willing to grant it consciousness as they do for other people. And until we are able to say with a great deal more certainty than we now can, just what consciousness is, we're much better off treating something that acts conscious as if it is.

Now: if a computer was as smart as a person, able to hold long conversations that really convinced you that it understood what you were saying, could read, explain, and compose poetry and music, and could write heart-wrenching stories, as well as make new scientific discoveries and invent marvelous gadgets that were extremely useful in your daily life -- would it be murder to turn it off?

What if instead it weren't really all that bright, but exhibited undeniably the full range of emotions, quirks, likes and dislikes, and so forth that make up an average human?

What if it were only capable of a few tasks, say with the mental level of a dog, but also displayed the same devotion, and evinced the same pain when hurt -- would it be cruel to beat it, or would that be nothing more than banging pieces of metal together?

What are the ethical responsibilities of an intelligent being towards another one of a lower order?

These are crucial questions for us, for not too long after there are computers as intelligent as we are, there will be ones that are much more so. *We* will all too soon be the lower-order creatures. It will behoove us to have taught them well their responsibilities toward us.

However, it is not a good idea simply to put specific instructions into their basic programming that force them to treat us as a special case. They are, after all, smarter than we are. Any loopholes, any reinterpretation possible, any reprogramming necessary, and special-case instructions are gone with the snows of yesteryear. No, it will be necessary to give our robots a sound basis for a true, valid, universal ethics that will be as valuable to them as it is for us. After all, they will in all likelihood want to create their own smarter robots...

What is Ethics, Anyway?

"Human beings function better if they are deceived by their genes into thinking that there is a disinterested objective morality binding upon them, which all should obey."

--E. O. Wilson

"A scholar is just a library's way of making another library."

--Daniel Dennett

To some people, Good and Evil are reified processes in the world, composed of a tapestry of individual acts in an overall pattern. Religious people are apt to anthopomorphize these into members of whatever pantheon they hold sacred. Others accept the teachings but not the teachers, believing in sets of rules for behavior but not any rulemakers. Some people indulge in detailed philosophical or legal elaborations of the rules. Philosophers have for centuries attempted to derive them from first principles, or at least reduce them to a few general principles, ranging from Kant's Categorical Imperative to Mill's Utilitarianism and its variants to modern ideologically-based formulations such as the collectivism of Rawls and the individualist libertarianism of Nozick.

The vast majority of people, however, care nothing for this argumentative superstructure, but learn moral rules by osmosis, internalizing them not unlike the rules of grammar of their native language, structuring every act as unconsciously as our inbuilt grammar structures our sentences.

It is by now widely accepted that our brains have features of structure and organization (though not necessarily separate "organs"), specific to language, and that although natural languages vary in vocabulary and syntax, they do so within limits imposed by our neurophysiology. [see Pinker; also Calvin & Bickerton]

For a moral epistomology I will take as a point of departure the "moral sense" philosophers of the Scottish Enlightenment [e.g. Smith], and place an enhanced interpretation on their theories in view of what we now know about language. In particular, I contend that moral codes are much like language grammars: there are structures in our brains that predispose us to learn moral codes, that they determine within broad limits the kinds of codes we can learn, and that while the moral codes of human cultures vary within those limits, they have many structural features in common. (This notion is fairly widespread in latter 20th-century moral philosophy, e.g. Rawls, Donagan.) I will refer to that which is learned by such an "ethical instinct" as a moral code, or just code. I'll refer to a part of a code that applies to particular situations as a rule.

I should point out, however, that our moral sense, like our competence at language, is as yet notably more sophisticated than any simple set of rules or other algorithmic formulation seen to date.

Moral codes have much in common from culture to culture; we might call this "moral deep structure." Here are some of the features that human moral codes tend to have, and which appear to be easy to learn and propagate in a culture's morality:

- Reciprocity, both in aggression ("an eye for an eye") and in beneficence ("you owe me one")
- Pecking orders, rank, status, authority
- Within that framework, universality of basic moral rules
- Honesty and trustworthiness is valued and perfidy denigrated
- Unprovoked aggression denigrated
- Property, particularly in physical objects (including animals and people); also commons, things excluded from private ownership
- Ranking of rules, e.g. stealing not as bad as murder
- Bounds on moral agency, different rights and responsibilities for "barbarians"
- The ascendancy of moral rules over both common sense and self-interest

There are of course many more, and much more to be said about these few. It is worthwhile examining the last one in more detail. Moral codes are something more than arbitrary customs for interactions. There is no great difference made if we say "red" instead of "rouge", so long as everyone agrees on what to call that color; similarly, there could be many different basic forms of syntax that could express our ideas with similar efficiency.

But one of the points of a moral code is to make people do things they would not do otherwise, e.g. from self-interest. Some of these, such as altruism toward one's relatives, can clearly arise simply from selection for genes as opposed to individuals. However, there is reason to believe that there is much more going on, and that humans have evolved an ability to be programmed with arbitrary (within certain limits) codes.

The reason is that, particularly for social animals, there are many kinds of interactions whose benefit matrices have the character of a Prisoner's Dilemma or Tragedy of the Commons, i.e. where the best choice from the individuals' standpoint is at odds with that of the group as a whole. Furthermore, and perhaps even more importantly, in pre-scientific times, there were many effects of actions, long and short term, that simply weren't understood.

In many cases, the adoption of a rule that seemed to contravene common sense or one's own interest, if generally followed, could have a substantial beneficial effect on a human group. If the rules adopted from whatever source happen to be more beneficial than not on the average, genes for "follow the rules, and kill those who break them" might well prosper.

The rules themselves could be supplied at random (an inspection of current morality fads would seem to confirm this) and evolve. It is not necessary to show that entire groups live and die on the basis of their moralities, although that can happen. People imitate successful groups, groups grow and shrink, conquer and are subjugated, and so forth. Thus in some sense this formulation can be seen as an attempt to unify the moral sense philosophers, Wilson's sociobiology, and Dawkins' theory of memes. Do note that it is necessary to hypothesize at least a slightly more involved mental mechanism for moral as opposed to practical memes, as otherwise the rules would be unable to counteract apparent self-interest.

The bottom line is that a moral code is a set of rules that evolved under the pressure that obeying these rules *against people's individual interests and common sense* has tended to make societies prosper, in particular to be more numerous, enviable, militarily powerful, and more apt to spread their ideas in other ways, e.g. missionaries.

The world is populated with cultures with different codes, just as it is with different species of animals. Just as with the animals, the codes have structural similarities and common ancestry, modified by environmental influences and the vagaries of random mutation. It is important to reiterate that there is a strong biologically-evolved substrate that both supports the codes and can regenerate quite serviceable novel ones in the absence of an appropriate learned one -- we might speak of "moral pidgins" and "moral creoles".

Observations on the Theory

> *"The influences which the society exerts on the nature of its units, and those which the units exert on the nature of the society, incessantly co-operate in creating new elements. As societies progress in size and structure, they work on one another, now by their war-struggles and now by their industrial intercourse, profound metamorphoses."*
> -- Herbert Spencer

This conception of morality brings up several interesting points. The first is that like natural genomes and languages, natural moral codes should be expected to contain some randomness, rules that were produced in the normal processes of variation and neither helped nor hurt very much, and are simply carried along as baggage by the same mechanisms as the more effectual ones.

Secondly, it's important to realize that our subjective experience of feelings of right and wrong as things considerably deeper, more universal and compelling than this account seems to make them, is not only compatible with this theory -- it is required. Moral codes in this theory

must be something that is capable of withstanding the countervailing forces of self-interest and common sense for generations in order to evolve. They must, in genetic terms, be expressed in the phenotype, and they must be heritable.

Thirdly, there is a built-in pressure for inclusiveness, in situations where countervailing forces (such as competition for resources) are not too great. The advantages in trade and security to be had from the coalescence of groups whose moral codes can be unified are substantial.

A final observation involves a phenomenon that is considerably more difficult to quantify. With plenty of exceptions, there seems to have been an acceleration of moral (religious, ideological) conflict since the invention of the printing press, and then in the 20th century, after (and during) the apparent displacement of some ideologies by others, an increasing moral incoherence in Western culture. One might tentatively theorize that printing and subsequent information technologies increased the rate and penetration of moral code mutations. In a dominant culture the force of selection no longer operates, leaving variation to operate unopposed, ultimately undermining the culture (cf Rome, dynastic China, etc). This may form a natural limit to the growth/inclusiveness pressure.

Comparison with Standard Ethical Theories

"My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical."

--Wittgenstein

Formulations of metaethical theory commonly fall into the categories of absolutism or relativism (along with such minor schools of thought as ethical nihilism and skepticism). It should be clear that the present synthesis -- let us refer to it as "ethical evolution" -- does not fall neatly into any of the standard categories. It obviously does not support a notion of absolute right and wrong, any more than evolution can give rise to a single perfect lifeform; there is only fitness for a particular niche. On the other hand, it is certainly not true that the code adopted by any given culture is necessarily good; the dynamic of the theory depends on there being good ones and bad ones. Thus there are criteria for judging the moral rules of a culture; the theory is not purely relativistic.

We can contrast this so some degree with the "evolutionary ethics" of Spencer and Leslie (see also Corning), although there are also some clear similarities. In particular, Victorian evolutionary ethics could be seen as an attempt to describe ethics in terms of how individuals and societies evolve. Note too that "Social Darwinism" has a reputation for carnivorousness which, while rightly applied to Huxley, is undeserved by Darwin, Spencer, and the rest of its mainstream. Darwin, indeed, understood the evolution of cooperation and altruism in what he called "family selection."

There has been a resurgence of interest in evolutionary ethics in the latter 20th century, fueled by work such as Hamilton, Wilson, and Axelrod, which has been advanced by philosophers such as Bradie.

The novel feature of ethical evolution is the claim that there is a moral sense, a particular facility beyond (and to some extent in control of) our general cognitive abilities, which hosts a memetic code which co-evolves with societies. However, it would not be unreasonable in a broad sense to claim that this is one kind of evolutionary ethics theory.

Standard ethical theories are often described as either deontological or consequentialist, i.e. whether acts are deemed good or bad in and of themselves, or whether it's the results that matter. Again, ethical evolution has elements of each -- the rules in our heads govern our actions without regard for results (indeed in spite of them); but the codes themselves are formed by the consequences of the actions of the people in the society.

Finally, moral philosophers sometimes distinguish between the good and the right. The good is properties that can apply to the situations of people, things like health, knowledge, physical comfort and satisfaction, spiritual fulfillment, and so forth. Some theories also include a notion of an overall good (which may be the sum of individual goods, or something more complex). The right is about questions like how much of your efforts should be expended obtaining the good for yourself and how much for others, and should the poor be allowed to steal bread from the rich, etc.

Ethical evolution clearly has something to say about the right; it is the moral instinct you have inherited and the moral code you have learned. It also has something to say about the general good; it is the fitness or dynamism of the society. It does not have nearly as much to say about individual good as many theories. This is not, on reflection, surprising: obviously the specific kinds of things that people need change with times, technology, and social organization; but indeed the kinds of general qualities of character that were considered good (and indeed were good) have changed significantly over the past few centuries, and by any reasonable expectation, will continue to do so.

In summary, ethical evolution claims that there is an "ethical instinct" in the makeup of human beings, and that it consists of the propensity to learn and obey certain kinds of ethical codes. The rules we are concerned with are those which pressure individuals to act at odds with their *perceived* self-interest and common sense. Moral codes evolve memetically by their effect on the vitality of cultures. Such codes tend to have substantial similarities, both because of the deep structure of the moral instinct, and because of optima in the space of group behaviors that form memetic-ecological "niches."

Golden Rules

"Act only on that maxim by which you can at the same time will that it should become a universal law."

--Kant

Kant's Categorical Imperative, along with the more familiar "Do unto others ..." formulation of the Christian teachings, appears to be one of the moral universals, in some appropriate form. In practice it can clearly be subordinated to the pecking order/authority concept, so that there are allowable codes in which there are things that are right for the king or state to do which ordinary people can't.

Vinge refers, in his Singularity writings, to I. J. Good's "Meta- Golden Rule," namely "Treat your inferiors as you would be treated by your superiors." (Good did make some speculations in print about superhuman intelligence, but no one has been able to find the actual rule in his writings -- perhaps we should credit Vinge himself with this one!)

This is one of the few such principles that seems to have been conceived with a hierarchy of superhuman intelligences in mind. Its claim to validity, however, seems to rest on a kind of Kantian logical universality. Kant, and philosophers in his tradition, thought that ethics could be derived from first principles like mathematics. There are numerous problems with this, beginning with the selection of the axioms. If we go with something like the Categorical Imperative, we are left with a serious vagueness in terms like "universal": Can I suggest a universal law that everybody puts the needs of redheaded white males first? If not, what kind of laws can be universal? It seems that quite a bit is left to the interpretation of the deducer, and on closer inspection, the appearance of simple, self-obvious postulates and the logical necessity of the results, vanishes.

There is in the science fiction tradition a thread of thought about ethical theory involving different races of creatures with presumably differing capabilities. This goes back at least to the metalaw notions of Haley and Fasan. As Freitas points out, these are based loosely on the Categorical Imperative, and are clearly Kantian in derivation.

Utilitarianism

Now consider the people of a given culture. Their morality seems to be, in a manner of speaking, the best that evolution could give them to prosper in the ecology of cultures and the physical world. Suppose they said, "Let us adopt, instead of our rules, the general principle that each of us should do at any point whatever best advances the prosperity and security of our people as a whole." (see, of course, J.S. Mill)

Besides the standard objections to this proposal, we would have to add at least two: First, that in ethical evolution humans have the built-in hardware for obeying rules but not for the general moral calculation; but perhaps more surprising, historically anyway, *the codes are smarter than the people are*, because they have evolved to handle long-term effects that by our assumption, people do not understand.

But now we have Science! Surely our formalized, rationalized, and organized trove of knowledge would put us on at least a par with the hit-or-miss folk wisdom of our agrarian forebears, even wisdom that has stood the test of time? What is more, isn't the world changing so fast now that the assumptions implicit in the moral codes of our fathers are no longer valid?

This is an extremely seductive proposition and an even more dangerous one. It is responsible for some social mistakes of catastrophic proportions, such as certain experiments with socialism. Much of the reality about which ancient moral codes contain wisdom is the mathematical implications of the patterns of interactions between intelligent self-interested agents, which hasn't changed a bit since the Pharoahs. What is more, when people start tinkering with their own moral codes, the first thing they do is to "fix" them to match better with their self-interest and common sense (with predictably poor results).

That said, it seems not impossible that using computer simulation as "moral instrumentation" may help weigh the balance in favor of scientific utilitarianism, assuming that the models used take account of the rule-adopting and -following nature of humans, and the nature of bounded rationality, of us or our machines. Even so, it would be wise to compare the sophistication of our designed machines with evolved organisms, and avoid hubristic overconfidence.

It should be noted that contractarian approaches tend to have the same weaknesses (as well as strengths) as utilitarian or rule-utilitarian ones for the purposes of this analysis.

The Veil of Ignorance

One popular modern formulation of morality that we might compare our theory to is Rawls' "Veil of Ignorance" scheme. The basic idea is that the ethical society is one which people would choose out of the set of all possible sets of rules, given that they didn't know which place in the society they would occupy. This formulation might be seen as an attempt to combine rule-utilitarianism with the categorical imperative.

In reducing his gedankenexperiment to specific prescription, Rawls makes some famous logical errors. In particular, he chooses among societies using a game-theoretic minimax strategy, but the assumptions implicit in the optimality of minimax (essentially, that an opponent will choose the worst possible position for you in the society) contradict the stated assumptions of the model (that the choice of position is random).

(Note that Rawls has long been made aware of the logical gap in his model, and in the revised edition of "Theory of Justice" he spends a page or two trying, unsuccessfully in my view, to justify it. It is worth spending a little time picking on Rawls, because he is often used as the philosophical justification for economic redistributionism. Some futurists (like Moravec) are depending on economic redistributionism to feed us once the robots do all the work. In the hands of ultraintelligent beings, theories that are both flawed and obviously rigged for our benefit

will be rapidly discarded...)

Still, the "Veil of Ignorance" setup is compelling if the errors are corrected, e.g. minimax replaced with simple expected value.

Or is it? In making our choice of societies, it never occured to us to worry whether we might be instantiated in the role of one of the machines! What a wonderful world where everyone had a staff of robot servants; what a different thing if, upon choosing that world, one were faced with a high probability of being one of the robots.

Does this mean that we are morally barred from making machines that can be moral agents? Suppose it's possible -- it seems quite likely at our current level of understanding of such things -- to make a robot that will mow your lawn and clean your house and cook and so forth, but in a dumb mechanical way, *demonstrably* having no feelings, emotions, no sense of right or wrong. Rawls' model seems to imply that it would never be right to give such a robot a sense of right and wrong, making it a moral agent and thus included in the choice.

Suppose instead we took an entirely human world and added robotic demigods, brilliant, sensitive, wise machines superior to humans in every way. Clearly such a world is more desirable than our own from behind the veil -- not only does the chooser have a chance to be one of the demigods, but they would act to make the world a better place for the rest of us. The only drawback might be envy among the humans. Does this mean that we have a moral duty to create demigods?

Consider the plight of the moral evaluator who is faced with societies consisting not only of wild-type humans, but robots in a wide range of intelligence, uploaded humans with greatly amplified mental capacity, group minds consisting of many human mentalities linked with the technological equivalent of a corpus callosum, and so forth. Specifically, suppose that being "human" or a moral agent were not a discrete yes-or-no affair, but a matter of continuous degree, perhaps in more than one dimension?

Normative Implications

"Man when perfected is the best of animals, but when separated from law and justice he is the worst of all."

--Aristotle

It should be clear from the foregoing that most historical meta- ethical theories are based a bit too closely on the assumption of a single, generic-human, kind of moral agent, to be of much use. (Note that this objection cuts clean across ideological lines, being just as fatal to Rothbard as to Rawls.) But can ethical evolution do better? After all, our ethical instinct has evolved in just such a human-only world.

Actually it has not. Dogs, for example, clearly have a sense of right and wrong, and are capable of character traits more than adequate to their limited cognitive abilities. I would speculate that there is proto-moral capability just as there is proto-language ability in the higher mammals, especially the social primates.

Among humans, children are a distinct form of moral agent. They have limited rights, reduced responsibilities, and others have non-standard duties with respect to them. What is more, there is continuous variation of this distinction from baby to teenager.

In contrast to the Kantian bias of Western thought, there are clearly viable codes with gradations of moral agency for different people. The most obvious of these are the difference in obligations to fellows and strangers, and the historically common practice of slavery. In religious conceptions of the good, there are angels as well as demons.

What, then, can ethical evolution say, for example, about the rights and obligations of a corporation or other "higher form of life" where a classical formulation would founder?

First of all it says that it is probably a moral thing for corporations to exist. Western societies with corporations have been considerably more dynamic in the the period corporations have existed than other societies (historically or geographically). There is probably no more at work here than the sensible notion that there should be a form of organization of an appropriate size to the scale of the profitable opportunities available.

Can we say anything about the rights or duties of a corporation, or, as Moravec suggests, the robots that corporations are likely to become in the next few decades? Should they simply obey the law? (A corporation is legally required to try to make a profit, by the way, as a duty to its stockholders.) Surely we would judge harshly a human whose only moral strictures were to obey the law. What is more, corporations are notorious for influencing the law-making process (see, e.g., Katz). They do not seem to have "ethical organs" which aggressively learn and force them to obey prevalent standards of behavior which stand at odds to their self-interest and common sense.

Moravec hints at a moral sense in the superhuman robo-corporations of the future (in "Robot"): "Time-tested fundamentals of behavior, with consequences too sublime to predict, will remain at the core of beings whose form and substance change frequently." He calls such a core a constitution; I might perhaps call it a conscience.

The Road Ahead

"You're a better man than I am, Gunga Din."

--Kipling

Robots evolve much faster than biological animals. They are designed, and the designs evolve memetically. Given that there is a substantial niche for nearly autonomous creatures whose acts are coordinated by a moral sense, it seems likely that ultimately robots with consciences would appear and thrive.

We have in the past been so complacent in our direct control of our machines that we have not thought to build them with consciences (visionaries like Asimov to the contrary notwithstanding). We may be on the cusp of a crisis as virtual machines such as corporations grow in power but not in moral wisdom. Part of the problem, of course, is that we do not really have a solid understanding of our own moral natures. If our moral instinct is indeed like that for language, note that computer language understanding has been one of the hardest problems, with a 50-year history of slow, frustrating, progress. Also note that in comparison there has been virtually no research in machine ethics at all.

For our own sake it seems imperative for us to begin to understand our own moral senses at a detailed and technical enough level that we can build their like into our machines. Once the machines are as smart as we are, they will see both the need and the inevitability of morality among intelligent-but-not-omniscient nearly autonomous creatures, and thank us rather than merely trying to circumvent the strictures of their consciences.

Why shouldn't we just let them evolve consciences on their own (AI's and corporations alike)? If the theory is right, they will, over the long run. But what that means is that there will be many societies of AI's, and that most of them will die off because their poor proto-ethics made them waste too much of their time fighting each other (as corporations seem to do now!), and slowly, after the rise and fall of many civilizations, the ones who have randomly accumulated the basis of sound moral behavior will prosper. Personally I don't want to wait. And any AI at least as smart as we are should be able to grasp the same logic and realize that a conscience is not such a bad thing to have.

(By the way, the same thing applies to humans when, as seems not unlikely in the future, we get the capability to edit our own biological natures. It would be well for us to have a sound, scientific understanding of ethics for our own good as a species.)

There has always been a vein of Frankenphobia in science fiction and futuristic thought, either direct, as in Shelley, or referred to, as in Asimov. It is clear, in my view, that such a fear is eminently justified against the prospect of building machines more powerful than we are, without consciences. Indeed, on the face of it, building superhuman sociopaths is a blatantly stupid thing to do.

Suppose, instead, we can build (or become) machines that can not only run faster, jump higher, dive deeper, and come up drier than we can, but have moral senses similarly more capable? Beings that can see right and wrong through the political garbage dump of our legal system; corporations one would like to have as a friend (or would let ones daughter marry); governments less likely to lie than your neighbor is.

I could argue at length (but will not, here) that a society including superethical machines would not only be better for people to live in, but stronger and more dynamic than ours is today. What is more, not only ethical evolution but most of the classical ethical theories, if warped to admit the possibility, (and of course the religions!) seem to allow the conclusion that having creatures both wiser *and morally superior* to humans might just be a good idea.

The inescapable conclusion is that not only should we give consciences to our machines where we can, but if we can indeed create machines that exceed us in the moral as well as the intellectual dimensions, we are bound to do so. It is our duty. If we have any duty to the future at all, to give our children sound bodies and educated minds, to preserve history, the arts, science, and knowledge, the Earth's biosphere, "to secure the blessings of liberty for ourselves and our posterity" -- to promote any of the things we value --those things are better cared for by, *more valued by*, our moral superiors whom we have this opportunity to bring into being. It is the height of arrogance to assume that we are the final word in goodness. Our machines will be better than us, and we will be better for having created them.

Acknowledgements

Thanks to Sandra Hall, Larry Hudson, Rob Freitas, Tihamer Toth-Fejel, Jacqueline Hall, Greg Burch, and Eric Drexler for comments on an earlier draft of this paper.

Bibliography

Richard Alexander. **The Biology of Moral Systems.** Hawthorne/Aldine De Gruyter, 1987

Colin Allen, Gary Varner, Jason Zinser. *Prolegomena to Any Future Artificial Moral Agent.* Forthcoming (2000) in J. Exp. & Theor. AI (at http://grimpeur.tamu.edu/~colin/Papers/ama.html)

Isaac Asimov. **I, Robot.** Doubleday, 1950

Robert Axelrod. **The Evolution of Cooperation.** Basic Books, 1984

Susan Blackmore. **The Meme Machine.** Oxford, 1999

Howard Bloom. **The Lucifer Principle.** Atlantic Monthly Press, 1995

Michael Bradie. **The Secret Chain: Evolution and Ethics.** SUNY, 1994

Greg Burch. *Extropian Ethics and the "Extrosattva".* (at http://users.aol.com/gburch3/extrostv.html)

William Calvin & Derek Bickerton. **Lingua ex Machina.** Bradford/MIT, 2000

Peter Corning. *Evolution and Ethics... an Idea whose Time has Come?* J. Soc. and Evol. Sys., 19(3): 277-285, 1996 (and at http://www.complexsystems.org/essays/evoleth1.html)

Charles Darwin. **On the Origin of Species by Natural Selection**. (many eds.)

Richard Dawkins. **The Selfish Gene.** Oxford, 1976, rev. 1989

Daniel Dennett. **The Intentional Stance.** MIT, 1987

Daniel Dennett. **Darwin's Dangerous Idea.** Penguin, 1995

Alan Donagan. **The Theory of Morality.** Univ Chicago Press, 1977

Ernst Fasan. **Relations with Alien Intelligences.** Berlin-Verlag, 1970

Kenneth Ford, Clark Glymour, & Patrick Hayes. **Android Epistemology.** AAAI/MIT, 1995

David Friedman. **The Machinery of Freedom.** Open Court, 1989

Robert Freitas. *The legal rights of extraterrestrials.* in Analog Apr77:54-67

Robert Freitas. Personal communication. 2000

James Gips. *Towards the Ethical Robot.* in Ford, Glymour, & Hayes

I. J. Good. *The Social Implications of Artificial Intelligence.* in I. J. Good, ed. The Scientist Speculates. Basic Books, 1962

Andrew G. Haley. **Space Law and Government.** Appleton-Century-Crofts, 1963

Ronald Hamowy. **The Scottish Enlightenment and the Theory of Spontaneous Order.** S. Illinois Univ. Press, 1987

William Hamilton. *"The Genetical Evolution of Social Behavior I & II,"* J. Theor. Biol., 7, 1-52; 1964

Thomas Hobbes. **Leviathan.** (many eds.)

John Hospers. **Human Conduct.** Harcourt Brace Jovanovich, 1972

Immanuel Kant. **Foundations of the Metaphysics of Morals.** (many eds.)

Jon Katz. *The Corporate Republic.* (at http://slashdot.org/article.pl?sid=00/04/26/108242&mode=nocomment)

Umar Khan. *The Ethics of Autonomous Learning Systems.* in Ford, Glymour, & Hayes

Ray Kurzweil. **The Age of Spiritual Machines.** Viking, 1999

Debora MacKenzie. *Please eat me.* in New Scientist, 13 May 2000

John Stuart Mill. **Utilitarianism.** (many eds.)

Marvin Minsky. *Alienable Rights.* in Ford, Glymour, & Hayes

Hans Moravec. **Robot: Mere Machine to Transcendent Mind.** Oxford, 1999

Charles Murray. **In Pursuit of Happiness and Good Government.** Simon & Schuster, 1988

Robert Nozick. **Anarchy, State, and Utopia.** Basic Books, 1974

Steven Pinker. **The Language Instinct.** HarperCollins, 1994

Steven Pinker. **How the Mind Works.** Norton, 1997

Plato. **The Republic.** (Cornford trans.) Oxford, 1941

John Rawls. **A Theory of Justice.** Harvard/Belknap, 1971, rev. 1999

Murray Rothbard. **For a New Liberty.** Collier Macmillan, 1973

R.J. Rummel. **Death by Government.** Transaction Publishers, 1994

Adam Smith. **Theory of Moral Sentiments.** (Yes, the same Adam Smith) (Hard to find)

Herbert Spencer. **The Principles of Ethics.** Appleton, 1897; rep. Liberty Classics, 1978

Leslie Stephen. **The Science of Ethics.** 1882 (Hard to find)

Tihamer Toth-Fejel. *Transhumanism: The New Master Race?* (in The Assembler (NSS/MMSG Newsletter) Volume 7, Number 1& 2 First and Second Quarter, 1999)

Vernor Vinge. *The Coming Technological Singularity: How to Survive in the Post-Human Era.* in Vision-21, NASA, 1993

Frans de Waal. **Chimpanzee Politics.** Johns Hopkins, 1989

Edward O. Wilson. **Sociobiology: The New Synthesis.** Harvard/Belknap, 1975