After the Humans are Gone

Eric Dietrich
Philosophy Dept.
Binghamton University
Binghamton NY 13902-6000
http://bingweb.binghamton.edu/~dietrich

Recently, on the History Channel, artificial intelligence (AI) was singled out, with much wringing of hands, as one of the seven possible causes of the end of human life on Earth. I argue that the wringing of hands is quite inappropriate: the best thing that could happen to humans, and the rest of life of on planet Earth, would be for us to develop intelligent machines and then usher in our own extinction.

1. Humans versus the world

British astrophysicist, Stephen Hawking, recently asked the following question on Yahoo Answers: "In a world that is in chaos politically, socially, and environmentally, how can the human race sustain another 100 years?" Some of the answers included: "Get rid of nuclear weapons," and "Somehow we will." A number of people suggested thinking differently: ending bickering or fostering cooperation. Many were doubtful that we could survive another 100 years.

What is the prognosis for the human race? In the long run, extinction: 99.9% of all plants and animals that have ever lived are now extinct. While it is true that we differ from all the other species in one important way (our intelligence), we are nevertheless a species quite similar to all the rest. So, simple induction implies that humans will one day go extinct. And this is true if nothing devastating happens. But something devastating is happening.

The background extinction rate is estimated at 2 – 4 families per million years. But this background extinction rate is swamped by mass extinctions. Paleontologists list five major mass extinctions over the last 600 million years: the Cretaceous-Tertiary, the End Triassic, the Permian-Triassic, the Late Devonian, and Ordovician-Silurian. Of course, most experts reckon that chances are tolerably low for an external major extinction event, at least in the immediate future . . . if you exclude humans.

Among the new things humankind brings to the world table is that *we ourselves are an extinction event*. Many biologists believe that we are currently in the early stages of a human-caused mass extinction, known as the *Holocene extinction event*. These biologists think that up to 20 percent of all living species could become extinct within 30 years. One-third of amphibians are at risk in the next few years. Biologist E.O. Wilson estimated in his 2002 book *The Future of Life* that if current rates of human destruction of the biosphere continue, one-half of all species of life on earth will be extinct in 100 years. So, humans *are* asteroids.

Given how devastating we are to the planet, and how entrenched our behavior is, an moral argument can be made that we *ought* to extinguish ourselves – and soon.

2. Humans versus humans.

In the last section, we saw that humans are bad for all the other living things on the planet. We are also bad for each other, because we are bad *to* each other. It is possible to survey humankind and be proud, for we accomplish great things. Art and science are two notable worthy human accomplishments. Consonant with art and science are some of the ways we treat each other. Sacrifice and heroism are two admirable human qualities that pervade human interaction. But, all this goodness is more than balanced by human depravity. Moral corruption infests our being. Why?

Throughout history, distinguished philosophers, theologians, and psychologists have wrestled with this question. Why are we so bad?

The Evolutionary basis of some immorality.

Let's focus on the badness that ordinary humans create while behaving more or less normally. By "normally," I mean that the behaviors I will consider are statistically common; they fall within the bump of the bell curve of human behaviors. I include in this set behaviors such as lying, cheating, stealing, raping, murdering, assaulting, mugging, child abuse, as well as such things as ruining the careers of, and discriminating against on the basis of sex, race, religion, sexual preference, and national origin.

How could ordinary humans have normal behavior that includes such things as rape, child abuse, murder, sexism, and racism? The standard answer locates the problem in us, but in such a way that mere moral discipline, perhaps enhanced by education, could fix it. For example, many claim that such bad behaviors are either learned or that the perpetrators have not learned ways of coping with the frustrations and aggravated selfishness that cause or lead to the bad behavior. Unfortunately, this answer is wrong. The correct answer is that many ordinary humans' worse behavior has an evolutionary explanation.

Consider two cases: child abuse and rape (for a more in-depth analysis which includes sexism and racism, see Dietrich 2001).

Child abuse

Here is a surprising statistic: the best predictor of whether or not a child will be abused or killed is whether or not he or she has a step-father (Wilson and Daly, 1988). Why should this be the case? Learning or lack of learning doesn't seem to be a plausible explanation here. Evolutionary theory, however, seems to succeed where the folk theory cannot. In some male-dominated, primate species (e.g., langurs), when a new alpha male takes over the troop, he kills all the infants fathered by the previous alpha male. He then mates with the females in his new harem, inseminating many of them, and now they will bear his children. The langur pattern is just one extreme case of a nearly ubiquitous mammalian phenomenon: males kill or refuse to care for infants that they conclude are unlikely to be their offspring, basing their conclusion on proximate cues. We carry this evolutionary baggage around with us.

Rape

The common explanation of rape is that it is principally about violence against women. The main consequence of this view is that rape is not sex. Many embrace this explanation simply because, emotionally, it seems right. But it is wrong (see, e.g., Thornhill and Palmer, 2000). Most rape victims around the world are females between the ages of 16 and 22, among the prime reproductive years for females (the best reproductive years are 19-24 or so, the overlap isn't exact). Most rapists are in their teens through their early twenties, the age of maximum male sexual motivation. Few rape victims experience severe, lasting physical injuries. On the available evidence, young women tend to resist rape more than older women. Rape is also ubiquitous in human cultures, there are no societies where rape is non-existent. Rape also exists in other most other animals: in insects, birds, reptiles, amphibians, marine mammals and non-human primates. All of these facts cry out for an evolutionary explanation of rape: rape is either an adaptation or a by-product of adaptations for mating. Either way, rape is part of the human blue-print.

So, on the best available theory we've got, two very serious social ills – child abuse and rape – are due to our evolutionary heritage (as are several other social ills). It is a sad fact that much of our basic, human psychological is built by evolution. These innate psychological capacities of ours are principally responsible for many of humanity's darkest ills. In short, we abuse, discriminate, and rape because we are human.

3. A modest proposal: *Homo sapiens* 2.0.

What can we do about the immorality humans perpetrate on each other and the thoughtless damage we do to the rest of the planet? The standard line taken by is to simply try to educate everyone to do better – to change society. But if the current evolutionary theories about some of our most dark behaviors are correct, such teaching either will not work, or will require draconian social measures. Yet, to those who think that producing better humans through teaching is a live option, I say: "Great – give it a try, what have you got to lose?" But I believe this path won't work. Suppose we try a better path.

Humankind shouldn't just go extinct. There are things about us worth preserving: art and science, to name two. Some might think that these good parts of humanity justify our continued existence. This conclusion no doubt used to be warranted, before AI became a real possibility. But now, it no longer is. If we could implement in machines the better angels of our nature, then morally we should, and then we should exit, stage left.

So, let's build a race of machines – *Homo sapiens* 2.0 -- that implement only what is good about humanity, that do not feel any evolutionary tug to commit certain evils against others of their own kind, and that let the rest of the world live in peace. And then let us – the humans – exit the stage, leaving behind a planet populated with nice machines, who, while not perfect angels, will nevertheless be a vast improvement over us.

One way to do this project would be to implement in the machines our best moral theories in such a way that the machines do not draw invidious distinctions. These are the theories that see morality as comprising universal truths, applying fairly to all beings. One such truth is that it is wrong to harm another being, normally. (I say "normally" because even in a better, machine society, it is likely there will be bad machines, and these must be dealt with.)

What are the prospects for building such a race of robots? They seem modestly high to me. The theories and technologies for building a human-level robot seriously elude us at the present time, but we have, I think, the correct foundational theory – computationalism (I have argued for this many times in various places; see Dietrich, 1994, Dietrich and Markman, 2000). Assuming that computationalism is correct, then it is only a matter of time before we figure out what algorithms govern the human mind. Once we know this, we could, with careful diligence, remove at least some of the parts responsible for behaving abominably. After building such a race of machines, perhaps we could exit with some dignity -- with the thought that we had finally done the best we could do.

4. An objection to *Homo sapiens* 2.0: Weinberg's Problem

I have received several objections to my proposal over the last few years. None work. Here, I want to rebut a new objection.

As mentioned, we should design the machines so they do not draw invidious distinctions, for these distinctions lie at the heart of immorality. The machines view themselves and all rest of the life on planet Earth with equal favor. The best way to accomplish this is to implement the machines as thorough-going scientific materialists. But, so the objection goes, the consequences of this are severe. It is not that the machines are merely scientific materialists, but rather the special way in which they became scientific materialists. The objection argues that the machines have a special epistemic status, and because of this status, nothing will awe or impress them; they will have no moral or spiritual fire to guide and inspire them. Lacking this, they will neither wonder nor explore, hence they will not create art or science. They will wind up being moral engineers -- perhaps building better and better versions of themselves, keeping this up until they have engineered a race of Buddhas, at which point they might reasonably stop. But such a world, the objection concludes, is worse than our current world, since it lacks inspiration and wonder, art and science. So we shouldn't build our machine replacements.

This objection is *not* the claim that because they are machines, our replacements will lack awe and wonder. The objection grants that the machines will have the capacity for full, inner lives, cognitively, emotionally, and phenomenologically. They will have desires, concerns, hope, cares, and beliefs. The problem is their special epistemic status. In our effort to keep from them from drawing invidious distinctions, we will see to it that they will inherit from us a purely scientific worldview -- a world of reasons and causes, laws and probabilities. The machines' worldview is therefore *rootless*: it is not rooted, as ours is, in awe and mystery, in reverence and wonder. Their scientific worldview is not hard-won, it is a gift. The sun will never be Helios or Ra to them; it's a large fusion reaction. Thunder is not the mighty Thor striking his magic hammer, Mjölnir, it is an acoustic shock wave caused by lightning rapidly heating and hence expanding the air. Love won't make their world go 'round; inertia will, and "make" will have to be written in scare-quotes. The machines will know who their creators were, and how flawed they (we) were. They won't be in awe of us; they may pity us while regarding us with some appreciation, since we (finally) did the right thing. The machines' existence won't even strike them as a fluke, as ours does now (to many). Instead, it will seem to them to be the next logical step. They will see themselves exactly as I have argued that they are – the rational, best alternative.

With such a hard-nosed view of their world and their place in it, the machines won't feel any angst, nor awe and wonder. And, lacking these states (remember, it is not that they *can't* feel awe and wonder, it is that they *don't*), they will not be driven to do art and science. They will not take risks. Since they can't be cowards, they won't be heroes. Something incalculably important will be lost, therefore, if we replace ourselves with the machines. No matter how good they are, no matter how much better for the other life on planet Earth, if we engineer these creatures and then embrace our own extinction, we will be extinguishing something profound, beautiful, and important.

What makes this objection interesting and powerful is that it is really a version of what I call *Weinberg's Problem*. In the closing lines of his 1977 book *The First Three Minutes*, the physicist, Stephen Weinberg, famously said: "The more the universe seems comprehensible, the more it also seems pointless." Being pointless and being unable to produce awe and wonder go hand in hand. Weinberg's Problem is *our* problem, of course, but the machines will have it in spades because of their rootless scientific materialism, which makes it far more thoroughgoing than ours.

I'm not saying their science won't have unanswered questions. They will inherit our science and ours is crawling with questions. That's not the issue. The issue is the world-view involved. In our noble effort to give them only what is best about us, and to not give them the wherewithal to do bad or evil acts either to the rest of life on Earth or to each other, we will be constrained to offer only what is rational, what is known, what can be counted on. The machines won't understand everything that happens, but they will think that everything that happens, happens either for some reason (using a variant of Leibnitz's "Principle of Sufficient Reason") or because of the relevant statistics, which is a kind of reason. Nor will they have an answer to every question. But because of their worldview, they will either dismiss such questions, or patiently seek to answer them. They will never experience majesty and grandeur in the world of ideas because none of the remaining scientific problems they have to solve will strike them as *deep*. They won't have any sort of spiritual, mysterious sense of what is deep. They will merely note that some problems are harder than others, and some, when solved, lead to solutions of many other problems. This is the extent of their notion of 'deep'.

So, lacking any sense of grandeur of their view of life and the world, they create no art and no profound science. They while away their lives being good and being good stewards. Yet this hardly seems to be enough for us to commit species-cide.

5. Reply – Attacking Weinberg's Problem head-on

There are several things to say to this Weinbergian objection. There are the whiny things to complain about: A) It assumes too tight a connection between being scientific materialists and lacking awe and wonder; B) It assumes too tight a connection between feeling awe and wonder and experiencing meaning; C) It assumes too tight a connection between being inspired by awe and wonder and doing science and art. But the very fact that Weinberg's Problem is an increasing problem *for us*, as our science advances, indicates that scientific materialism conflicts with meaningfulness and with being awed and inspired. The machines are more ensnared in Weinberg's Problem because of the rootless nature of their knowledge and worldview. But we will one day be as ensnared as they. Whether we replace ourselves by the machines or not, Weinberg's Problem looms on the horizon for any of Earth's resident intelligent entities.

The best way to attack Weinberg's problem is head-on. It isn't true that none of the scientific or mathematical problems they work on will strike them as deep. They cannot avoid developing a stance of profundity toward the universe they will inhabit. The machines perhaps won't marvel at a sunrise (what they will call an "earth-rotate"), but the universe is filled with other things that that they can marvel at. There are rock solid facts in our world that are positively shocking, and these facts are fully capable of inspiring awe and wonder, even if one is a hard-bitten scientific materialist. In fact, we ourselves have actually been doing a good job of ignoring these facts, but I think it is time to face up to them.

Many of these problems are actually well-known. They are the problems of *philosophy*. Why does dualism seem true? Why is consciousness impossible to reductively explain? Why are there subjective points of view? Where does our sense of self and freewill come from? Why is it so strongly felt but vanishes when science goes looking for it? What is the nature of being? . . . of morality?

It is not so much the specifics of philosophy's problems, it is their intractability, their immortality that is puzzling. Here it is, early in the twenty-first century, and Aristotle and Plato are still our colleagues. In no other field is this true. Aristotle, a genius polymath, is not today the colleague of any biologist, physicist, nor geologist -- in these areas, his theories were very wrong – not even in the ballpark. But in philosophy, if his office were down the hall, we'd go talk to him regularly. Our replacement machines will know this, since they will know the history of our philosophy.

They will also be conscious. And their consciousnesses will also strike them as not logically supervenient on the physical. Yet, they might well suppose that it is, like we do. They will be therefore be stuck with complete inexplicability of consciousness.

The machines will be far more moral than we. But they won't know the answer to this question: is the moral a function of ends, or is it inherent in an action, a deed? Like Aristotle, both Kant and Mill are still our colleagues, and they will be the machines' as well.

Finally, the machines will also have and be able to switch between subjective and objective points of view. And, this fact will be as paradoxical to them, as it is to us. Then, they will see a stark truth: switching between the subjective and objective creates the very problems of philosophy they grapple with (for more on this, see Nagel, 1986).

Once seen, the machines will locate versions of this paradox in mathematics, logic, and physics. It is not too much to suppose that at this point, some of the machines will begin to wonder: "Why are all these problems so intractable? What's going on?" Such wondering can turn to wonder.

Pablo Picasso once said: "Computers are useless, all they can give you are answers." But Picasso was wrong. Our replacement machines will ponder deep questions -- questions that will cause them to wonder with awe at the nature of the universe and their place in it -- questions that cause them to become *philosophers*. And from there, everything is possible, except of course, answers.

References

Dietrich, E. (ed.) 1994. *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines*. San Diego: Academic Press.

Dietrich, E. 2001. Homo sapiens 2.0: Why we should build the better robots of our nature. *J. Exper. & Theor. AI*, 13 (4), 323-328.

Dietrich, E. and Markman, A. 2000. *Cognitive Dynamics*. Mahwah, NJ: Lawrence Erlbaum.

Nagel, T. 1986. *The View From Nowhere*. New York: Oxford.

Thornhill, R. and Palmer, C. 2000. *A Natural History of Rape*. MIT: Boston.

Weinberg, S., 1977. *The First Three Minutes*. Basic Books/Perseus.

Wilson, M. Daly, M. 1988. *Homicide*. Aldine de Gruyter.