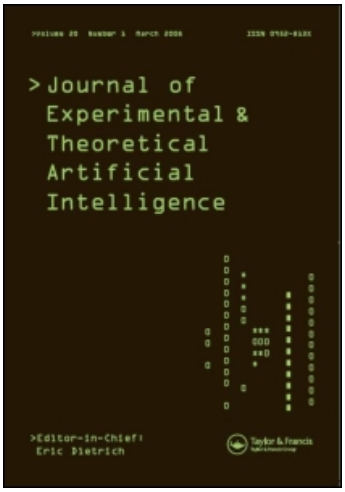

Journal of Experimental & Theoretical Artificial Intelligence
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713723652

After the humans are gone Douglas Engelbart Keynote Address, North American Computers and Philosophy Conference Rensselaer Polytechnic Institute, August, 2006

Eric Dietrich [a]
[a] Philosophy Department, Binghamton University, Binghamton, NY 13902-6000, USA

To cite this Article Dietrich, Eric(2007) 'After the humans are gone **Douglas Engelbart Keynote Address, North American Computers and Philosophy Conference Rensselaer Polytechnic Institute, August, 2006**', Journal of Experimental & Theoretical Artificial Intelligence, 19: 1, 55 — 67
To link to this Article: DOI: 10.1080/09528130601115339
URL: http://dx.doi.org/10.1080/09528130601115339

Journal of Experimental & Theoretical Artificial Intelligence
Vol. 19, No. 1, March 2007, 55–67

After the humans are gone
Douglas Engelbart Keynote Address, North American Computers and Philosophy Conference Rensselaer Polytechnic Institute, August, 2006

ERIC DIETRICH*

Philosophy Department, Binghamton University, Binghamton, NY 13902–6000, USA

(*Received 20 August 2006; in final form 15 November 2006*)

Artificial intelligence has long suffered the slings and arrows of humanists arguing in various ways that AI is dangerous to humanity. I argue the opposite: it is humanity that is dangerous, and replacing us by intelligent machines or agents would vastly improve the entire world. My argument beings with two observations. First, humans are extremely dangerous to all the other life on the planet. Second, humans are dangerous to all the other humans. Viewed objectively, getting rid of humans would be a good thing. Yet, it seems obvious that the good for which humans are responsible outweighs humans' negative effects. But what if we could replace humans with beings just as good, or better than we, yet with fewer negative effects? AI provides just this opportunity. I then consider one potent objection to this proposal. Because of their special epistemic status as engineered intelligences, the machines will lack the ability to be awed and inspired by their world. Lacking connections of wonder and inspiration to their world, they will lack the impetus necessary to create art and science. While their world might be better morally than ours, something of incalculable beauty will be lost if we turn Earth over to them. I show that this objection doesn't work.

Keywords: Human extinction; Moral corruption; Artificial intelligence; Robots

*Email: dietrich@binghamton.edu

Journal of Experimental & Theoretical Artificial Intelligence
ISSN 0952–813X print/ISSN 1362–3079 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/09528130601115339

1. Humans versus the world: why we will and should become extinct

The British astrophysicist, Stephen Hawking, recently asked the following question on Yahoo Answers (the site where anyone can pose a question for fellow Internet users): '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 (this estimate is from scientists at the American Museum of Natural History). Our species is called *Homo sapiens* (which means 'wise or knowing human (man)'). While it is true that we differ from all other species in one important way (our intelligence), we are nevertheless a species quite similar to all the rest. Therefore simple induction implies that one day humans will become extinct—and this is true if nothing devastating happens. However, something devastating could happen.

The background extinction rate is estimated at two to four families per million years (that is *families*, which are groups of groups of similar species (as in Kingdom, Phylum, Class, Order, Family, Genus, Species)), but this background extinction rate is swamped by mass extinctions. Paleontologists list five major mass extinctions over the last 600 million years.

- **Cretaceous–Tertiary extinction** occurred about 65 million years ago and was probably caused or aggravated by an impact of a several-mile-wide asteroid that created the Chicxulub crater now hidden off the Yucatan Peninsula and beneath the Gulf of Mexico (some argue for other causes, including gradual climate change or flood-like volcanic eruptions of basalt lava from India's Deccan Traps). The extinction killed 16% of marine families, 47% of marine genera, and 18% of land vertebrate families, including the dinosaurs.
- **End Triassic extinction** occurred roughly 199–214 million years ago, and was probably caused by massive floods of lava erupting from the central Atlantic magmatic province, an event that triggered the opening of the Atlantic Ocean. The volcanism may have led to deadly global warming. Rocks from the eruptions now are found in the eastern USA, eastern Brazil, North Africa, and Spain. The death toll was 22% of marine families and 52% of marine genera. The number of land vertebrate deaths remain unclear.
- **Permian–Triassic extinction** occurred about 251 million years ago. The Permian–Triassic catastrophe was Earth's worst mass extinction, killing *95% of all species* (53% of marine families, 84% of marine genera, and an estimated 70% of land species such as plants, insects, and vertebrates). Many scientists suspect a comet or asteroid impact, although direct conclusive evidence has not been found. Others believe the cause was flood volcanism from the Siberian Traps and related loss of oxygen in the oceans.
- **Late Devonian extinction** (cause unknown) occurred about 364 million years ago. It killed 22% of marine families and 57% of marine genera.
- **Ordovician–Silurian extinction** occurred about 439 million years ago, caused (possibly) by a drop in sea levels as glaciers formed, and then by rising sea levels

as glaciers melted. The toll was 25% of marine families and 60% of marine general.

(These facts are quoted from http://www.dimaggio.org/Evolution/5major.htm which, in turn, is taken from the work of paleobiologist Doug Erwin of the Smithsonian Institution's National Museum of Natural History. Estimates of extinction rates are from the late John J. Sepkoski, University of Chicago.)

Of course, most experts believe that chances are tolerably low for an external major extinction event, at least in the near future. *Tolerably low*? Here is a way to see the issue.

The average American's chances of dying as a result of an asteroid impact over a 50 year period is about the same as an average American's chances of dying in a tornado (approximately one in 20 000). Although many tornadoes occur each year, and some of them kill up to dozens of people, large asteroid impacts occur very rarely, but would kill millions or even billions of people. Consider this similar example. Several US citizens die every day, one at a time, by accidental electrocution; roughly the same number die on average in the very few jet airliner crashes each year, where hundreds die at one time. Airline crashes are relatively rare events: there were only two days during 2001 on which people died in airliner crashes in the USA; one, of course, was September 11. Asteroid impacts are just an extreme example of a rare but extraordinarily deadly event (Chapman and Morrison 1989).

So perhaps we are unlikely enough to become extinct because a comet or asteroid hits the Earth to relax a little. However, this is not the end of the story for human mass extinction. Unfortunately, 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 believe that up to 20% of all living species could become extinct within 20 years (by 2028). One-third of amphibians are at risk in the next few years. Biologist E.O. Wilson estimated that if current rates of human destruction of the biosphere continue, half of all species of life on earth will be extinct in 100 years (Wilson 2002). Humans *are* asteroids. We are a large mammal. We need a lot of food and water. Therefore it is not a big leap to conclude that we will be among the species wiped out.

The issue is far more pointed than this, however. Not only will humans become extinct eventually, but given how devastating we are to the planet, and how entrenched our behaviour is, an argument can be made that we *ought* to extinguish ourselves—and soon. The same conclusion is supported by an argument from another direction.

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 that we treat each other. Sacrifice and heroism are two

admirable human qualities that pervade human interaction. However, 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? How does one explain the Timothy McVeighs of the world? The Jeffrey Dahmers, the Ted Bundys? The Pol Pots, the Hitlers? The WTC terrorists? How are we to understand Charles Whitman (the University of Texas clock tower sniper) and Eric Harris and Dylan Klebold (the two Columbine School killers)? All these cases are baffling to the point of stupefaction. And we are powerless to prevent future monsters from killing us.

Immoralities that are less focused, which do not, as it were, have a point man, are equally bad, but more distributed. Sexism and racism, pervasive and damaging in the extreme, plague our lives. Of course, egregious cases of sexism and racism are often reported by individuals, and these are usually quite awful, but milder versions of sexism and racism probably inhabit each of us to some extent.

War is a horrible evil. Very few wars throughout history were what we might call 'just wars'. Wars are fought for greedy reasons—at least that is often why they start. War is also a persistent and common evil. Concerning the September 11 terrorist attacks in the USA, President Bush said: 'This is the beginning of the first war of the twenty-first century', as if it was inevitable there would be a first war of this century, followed by many more—and surely he was correct in that belief.

Then there are the horrors we live with each day: rape, murder, theft, assault, and the various new 'rages'—road rage, air rage, referee rage (admittedly not usually lethal, but always damaging; whoever said 'Sticks and stones may break my bones, but names will never hurt me' must have lived a solitary life on Mars).

Therefore we humans live out our lives suffering and causing harm great and small, eking out some measure of happiness via our art, our science, our loves, and our passions. Life is nasty, brutish, and, once in a while, beautiful . . . and long or short, depending on which part of it you happen to be experiencing.

2.1 *The evolutionary basis of some immorality*

Let us focus on the badness or evil that ordinary humans create while behaving more or less normally. By 'normally,' I mean that the behaviours I will consider are statistically common, that they fall within the hump of the bell curve of human behaviours. I include in this set behaviours such as lying, cheating, stealing, raping, murdering, assaulting, mugging, and child abuse, as well as such things as ruining the careers of, and discriminating against, people on the basis of sex, race, religion, sexual preference, and national origin. Not all of us have raped or murdered, but many of us have thought about it. Virtually all of us have lied, cheated, or stolen at some time in our lives. I intend to exclude war from my discussion, as well as such humans as Hitler, Pol Pot, Timothy McVeigh, the Columbine murderers, the recent hijacking terrorists, etc. Beings such as these are capable of extraordinary evil—evil that even if in some sense provoked (if only in the mind of the perpetrator), far outstrips the provocation. Beings such these commit gargantuan evil. I have no idea how to explain such beings, nor such evil. Like you, I can only shrug my shoulders

and point vaguely in the direction of broken minds working in collusion with random circumstances.

How could ordinary humans have normal behaviour which includes such things as rape, child abuse, murder, sexism, and racism? One standard answer is that such behaviours arise because of our innate selfishness, which can be overcome, at least in principle, by education or by a correct, happy upbringing (in all the cases of bad behaviour we will consider below, this standard answer is behind the scenes, working to supply energy to the incorrect folk explanation of the behaviours). This answer is wrong, at least for many of our immoral behaviours. The reasoning is simple. Selfishness alone cannot explain why we rape or kill our children: If we are all selfish but few of us murder or rape, then something else must be going on. The standard reply to this is that such bad behaviours are either learned or that the perpetrators have not developed ways of coping with the frustrations and aggravated selfishness that cause or lead to the bad behaviour. Unfortunately, this answer cannot be falsified, and moreover it does not explain some rather striking facts. The correct answer is that the bad behaviour of many ordinary humans has an evolutionary explanation, arising because we are animals that *evolved*, that have an evolutionary history dating back, through our immediate ancestors, almost 12 million years, and of course a continuous lineage dating back 3.5 billion years when life started on planet Earth. As Dennett (1995) has argued, all explanations of the way humans are must be grounded in the way we are made. And we are made by evolution. Let us explore the hypothesis that we are bad in part because of our evolutionary history by considering four cases: child abuse, sexism, rape, and racism.

2.1.1 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 stepfather. (The data suggest that abuse is meted out to older children; young children may be killed.) Why should this be the case? Learning or lack of learning does not seem to be a plausible explanation here. However, evolutionary theory 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.

2.1.2 Sexism. Our sexism is explained the same way. First, though, here is an interesting fact: every human culture is male dominated, and females are discriminated against in every culture. There are matrilineal cultures, but not female-dominated ones (the Amazons were a myth). What would explain this ubiquity of sexism? It obviously cannot be learned behaviour because the behaviours which we are certain are learned are not ubiquitous (e.g. driving on the left). Learned behaviours always vary substantially around the globe. Certainly, how men and women implement their inherent sexism is probably learned (e.g. always hold a door open for a woman, never let a woman vote), but discriminating against the female sex is not learned. It is part of our evolutionary

heritage—our evolutionary baggage. Why? Because we evolved from a male-dominated primate species (not all primate species are male dominated; some (vervets, many lemurs) are female dominated). In our cousin male-dominated species, it is males that typically get first helpings of the food, have the best locations for shelter, are groomed the most, etc. Females in these species frequently get seconds and the second-best in everything. Evolving from a species like this, human males naturally tend to think of human females as second-class members of the culture. (This explanation is a case of inference to the best explanation. We do not have access (or enough access) to the behaviours of the species we evolved from to say with complete conviction that we evolved from a male-dominated species. Nevertheless, this explanation is compelling in part because it best explains the ubiquity of sexism and it coheres best with what we know about other primate species.)

2.1.3 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. 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 approximately 19–24; the overlap is not exact). Most rapists are in their teens to 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 (interpretations of the anthropological findings of Turnbull and Mead are incorrect). Rape exists in other animals: insects, birds, reptiles, amphibians, marine mammals, and non-human primates. All these facts cry out for an evolutionary explanation of rape: rape is either an adaptation or a byproduct of adaptations for mating. Either way, rape is part of the human blueprint.

2.1.4 Racism. Although it is still somewhat disputatious, it is now reasonably clear that part of the engine of human evolution was group selection. Standard evolutionary theory posits that the unit of selection is the individual of a species. But selection pressures exist at many levels of life, from the gene level all way up to whole populations, communities, and even ecosystems—perhaps even to memes (roughly, culturally transmitted ideas). One such level is the group level, i.e. the level at which the traits of one member of a population affect the success of other members. It is known that group selection can produce species with properties that are not able to be evolved by individual selection alone (e.g. altruism). Group selection works by encouraging cooperation between members of the group and, often, discouraging cooperation between members of different groups. Therefore group selection has a dark side. Not only does it encourage within-group cooperation but, where groups overtly compete, it also tends to produce between-group animosity. Therefore, from our evolutionary past, humans tend to belong to groups, bond with the members of their own group, and fight with members of outlying groups. Which particular groups you feel compelled to hate (or dislike) is a

matter of historical accident and bad luck. But the fact that you tend to hate (or dislike) members of other groups is part of your genetic make-up.

To conclude, on the best available theory we possess, four very serious social ills—child abuse, sexism, rape, and racism—are due to our evolutionary heritage. It is a sad fact that much of our basic human psychology is built by evolution (and not by socialization (learning), as many believe, although, of course, socialization plays some role). These innate psychological capacities are principally responsible for many of humanity's darkest ills. In short, we abuse, discriminate, and rape because we are human. If we add on top of this that we also almost certainly lie, cheat, steal, and murder because we are human, we arrive at the idea that our humanity is the source of much of our anguish and suffering.

3. A modest proposal: *Homo sapiens* 2.0

The question naturally presents itself: 'What can we do about the immorality that humans perpetrate against each other and the thoughtless damage we do to the rest of the planet?' The standard line taken is simply to try to educate everyone to do better—to change society. However, if the current evolutionary theories about some of our darkest behaviours 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 do not believe that this path will work. Suppose we try a better path.

Humankind should not just become 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 us build a race of machines, *Homo sapiens* 2.0, which implement only what is good about humanity, which do not feel any evolutionary tug to commit certain evils, and which can let the rest of the world live. And then let us, the humans, exit the stage, leaving behind a planet populated with machines (or robots—all that really matters is that they are agents, and so I will just call them 'machines') which, while not perfect angels, will nevertheless be a vast improvement on us.

One way of carrying out this project would be to implement in the robots our best moral theories. These are the theories which see morality as comprising universal truths, applying fairly to all beings. One such truth is that it is normally wrong to harm another being. (I say 'normally' because, as I will discuss below, even in a better robot society, it is likely there will be bad robots, and these must be dealt with. Also, care must be taken here not to define 'harm' too narrowly. Dental work hurts, but it is not harming the individual.) Many of us, and many religions (but not all), aspire to such a morality. For example, Christians say 'Love thy neighbour', and sometimes, on their very best days, they define everyone, and sometimes every living thing, as their neighbour. Many Buddhists, at least in theory, aspire to such an inclusive definition.

What are the prospects for building such a race of robots? They seem modestly high to me. True, it is an immense leap from current technology to robustly intelligent machines. However, we are babes in the woods when it comes to AI and robotics; we are making decent advances and there is every reason to be optimistic. The theories and technologies for building a human-level robot seriously elude us at the present time, but we have, I believe, the correct foundational theory—computationalism (I have argued for this many times in various papers, and so I will spare you the arguments here). Assuming that computationalism is correct, then it is only a matter of time before we work out the algorithms that 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 that, we will be anachronistic—our presence will be at best unnecessary. 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 (I review the most common of them in the appendix). Here, I want to rebut a new and worrisome objection. The objection is as follows.

We should design the machines so they cannot draw invidious distinctions, for these distinctions lie at the heart of immorality. They view themselves and all rest of the life on planet Earth with equal favour. They do not think that they outrank most of the universe, nor do they think that some part of it outranks them. The best way to accomplish this is to implement the machines as thorough-going scientific materialists. However, this means that 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, continuing until they have engineered a race of Buddhas, at which point they might reasonably stop. However, such a world, the objection continues, is worse than our current world. Therefore we should not build our machine replacements.

What makes this objection interesting and difficult to handle is that it is *not* the claim that because they are machines, our replacements will lack awe and inspiration. 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. They will want to continue to exist. For example, they might well develop the technology needed to defend the Earth against collisions with asteroids. Rather, the objection insists that it is because of their *special epistemic status* that they lack awe, that they are not inspired as we are.

What is their special epistemic status? By assumption, the machines are mentally and psychologically quite similar to us, except that they are more intelligent and more moral. However, in our efforts to keep them from drawing invidious distinctions, we will see to it that they will inherit from us a purely scientific world-view—a world of reasons and causes, laws and probabilities. Therefore the machines' world-view is *rootless*; it is not rooted in awe and mystery, in reverence and wonder.

Their scientific world-view is not hard won: it is a gift. They will not have to work out that the world is not filled with gods and goddesses; such knowledge will be their default starting state. They will never live in a world made up of four elements. The sun will never be Helios or Ra to them. Thunder is not the mighty Thor striking his magic hammer Mjölnir; it is an acoustic shock wave caused by lightning rapidly heating and expanding the air. Love will not 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 will not be in awe of us; they may pity us while regarding us with some appreciation, since we did the right thing, for once. The machines' existence will not 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. They might feel sad that we had to die so they could live, but they might well see all naturally evolved life as inherently flawed both morally and epistemologically, and conclude that their existence is supremely justified.

In an effort to make the machines more moral than we are, we make them such hard-nosed materialists that they lose the distinction between the sacred and the profane. The machines will be rational scientific materialists, for this is what our best science is, and this is the best way to block drawing dangerous invidious distinctions. The machines will, in so far as possible, be objective. They will *calculate* the moral thing to do. And they will be coldly analytical about their inner lives. All this is required since it is the surest way to guarantee the morally best and brightest; it is the best path to a fair, just society of moral beings.

But any beings with such a hard-nosed view of their world, and their place in it, will not feel any angst, and hence no awe and wonder. Lacking these states (as a matter of fact, not as a matter of necessary design), they will not be driven to do art and science. They will not take risks. And, since they cannot be cowards, they will not be heroes. Therefore something incalculably important will be lost 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.

Remember, it is not that they *cannot* feel awe and wonder; it is that they *do not*. (I suppose we might build the machines so that they cannot feel awe and wonder. We might do this either because we conclude that it is actually a good thing to do or because lacking awe and wonder is a side-effect of other things we might implement in our machines. For example, we might build the machines so that they never invest themselves (or never invest themselves too heavily) in their possessions or their projects. Doing this seems like a reasonable way of preventing invidious distinctions, but feeling awe and wonder might require investing oneself in things. In any case, this is not the path to building *Homo sapiens* 2.0 asumed by the objection. Therefore I will set this to one side.)

This objection against my proposal is fundamentally a version of what I call *Weinberg's Problem*. In the closing lines of his book *The First Three Minutes* (Weinberg 1977, Basic Books/Perseus), the well-known physicist, Stephen Weinberg, famously said: 'The more the universe seems comprehensible, the more it also seems pointless'. Being pointless and lacking in awe and wonder go hand in hand. Weinberg's Problem is *our* problem, of course, but the machines will have it in

spades, since their world will be almost completely comprehensible right from the beginning.

I am not saying that their science will not have deep holes in it. They will inherit our science which is crawling with deep problems and puzzles. That is not the issue. The issue is the world-view involved. In our noble effort to give them only what is best about us, and not to 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 will not understand everything that happens, but they will think that everything that does happen, 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 world-view, they will be apt to dismiss such questions. 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 will not 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'. And this notion is not hardwired into them. Rather, it is acquired via the kind of beings they are and the kinds of lives they will lead.

Therefore, 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 seems to be not enough. Certainly it is not enough for us to commit species-cide.

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

There are several things to say about this Weinberg's Problem 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 being inspired by awe and wonder and doing science and art. Perhaps just curiosity and intelligence are all that is required, or curiosity, intelligence, and a sense of the beautiful, which by assumption, the machines would have. Or perhaps being only locally or narrowly fascinated is all that is required: 'Oh wow! Chlorophyll!' Perhaps. But the very fact that Weinberg's Problem is an increasing problem *for us*, as our science advances, indicates that scientific materialism and being awed and inspired are incompatible. The machines are more ensnared in Weinberg's Problem because of the rootless nature of their knowledge and worldview. However, one day we will be as ensnared as they, and we have deep epistemological roots. Whether we replace ourselves by the machines or not, Weinberg's Problem looms on the horizon for any of Earth's resident intelligent entities.

Perhaps the objection trades on a kind of nostalgia. A world filled with very moral kind engineers, tinkering away at various projects may be, in fact, a much better world that ours. We should just get over it, build the machines, and bow out. Well, maybe . . .

A much better way to attack this problem is head-on. The machines will not marvel at a sunrise (what they will call an 'earth-rotate'), but the universe is filled with other things that they can marvel at. There are rock solid facts in our world that are positively shocking. These are deeply puzzling and, I think, 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 puzzles, but I beleive that it is time to highlight 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 explain reductively? Why are there subjective points of view? Where does our sense of self and free will come from? Why is it so strongly felt, but vanishes when science goes looking for it? What is the nature of being?

It is not so much the specifics of philosophy's problems, but their intractability, their immortality, that is puzzling. Here we are, 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, or geologist; in these areas, his theories are very wrong—not even in the ballpark. But in philosophy, if his office were down the hall, we would go and 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. However, they might well suppose that it is, as we do. Therefore they will be stuck with the complete inexplicability of consciousness (see Dietrich and Hardcastle 2004).

The machines will also have subjective states and objective states. Just this brute fact will be puzzling to them, as it is to us. And its role in producing philosophy will not be lost on them. Among their number, there is bound to be a machine which, like the philosopher Thomas Nagel in one of the most famous papers in philosophy in the last 50 years, wonders what is like to be a bat (Nagel 1974). It is not too much to suppose that at least 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.

The machines will be far more moral than we are. But they will not know the answer to the 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' colleagues as well.

Mathematics and logic also provide a wealth of mysteries far more profound than mere puzzles. Are there true contradictions? If not, then why does it seem that there are? For example, why is mathematics bedeviled with paradoxes like those of Cantor, Burali-Forti, and Russell (respectively, the set of all sets is both the largest and not the largest set, there both is and is not a largest ordinal number (and that number is larger than itself), and there is a set which contains itself if and only if it does not contain itself)? Why does infinity come in sizes? It is not sufficient to know that it does. Rather, we need to know why this is, how it can be. Why is self-reference so productive of strange puzzles and truths like Gödel's theorem? In fact, why are there so many more mathematical truths than there are theorems to prove them?

What happens with limits (of which self-reference is a kind)? Consider the limits of what is conceivable or imaginable. Beyond these limits, there is the inconceivable or the unimaginable. Yet we have just conceived and imagined things beyond

these limits. How can that be? Whatever the mind is, it does appear to be very mysterious. Eventually, this mystery will occur to our replacement machines.

Pablo Picasso once said: 'Computers are useless; all they can give you are answers'. But this is not true. When we have finished building our replacement machines, they will be useful, even to Picasso, for they will *ask deep questions*, the very questions we ask—questions that 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.

Appendix: Some objections to building *Homo sapiens* 2.0

The most common objection to my proposal is that the robots will have their own evil behaviour. For one thing, we will have to program in self-preservation. And, as is well-known from AI, most problem solving takes place under time-pressure and with imperfect data. Therefore mistakes are likely to ensue. For example, it is likely that eventually a robot or group of robots will erroneously conclude that their lives are in some sort of danger from another robot or robot group and react accordingly, harming innocent robots.

Yes, probably this would happen. Probably the robots I am advocating would have their own suite of bad behaviours. But even if we could not eliminate all evil and harm, we should still eliminate what we can, just on standard consequentialist moral arguments. And eliminating everything from abuse through murder to discrimination and rudeness is eliminating quite a lot.

Another objection is that we cannot eliminate emotions like envy, jealousy, and rage without also eliminating all the good emotions like love, caring, and sympathy. This is a worrisome objection because good and evil might be two sides of the same coin, or different arcs of the same circle. However, we are ignorant enough of how emotions work and why they evolved to take seriously the idea that it is quite possible to have only good emotions. After all, many conceive of Heaven as just such a place—a place where there are no negative emotions, not even sadness. (I am not imagining that our robots will not be sad.) All I am suggesting is that we plausibly have the power to implement Heaven on Earth by implementing very moral robots.

Am I suggesting that we eliminate emotions altogether? I am not. However, it is not obvious that this is a bad idea, assuming of course it is even possible, for certain cognitive activity may, for all we know now, require certain emotions. Here, I am not just referring to our cognitive activity of thinking about our emotions. It may be that one cannot do science without loving knowledge or curiosity or something of this sort.

A third objection is not to build the robots, but to change humans via genetic engineering so that they commit either no evil or much less evil. However, I doubt that the required cognitive and emotional changes to 'edit' out rape, murder, etc. can be accomplished just by altering our genome. Fundamental changes must be made in how the world and its inhabitants are conceived. Genetically altered humanoids are unlikely to be able to implement those changes. Creatures, beings who could not

discriminate, could not rape, could not murder, would not be human, and, I am suggesting, not even animal.

A couple of quick caveats. It is a virtual certainty that robots will not have sexes or mate as we do. This, the cynic might say, already makes them way ahead of us in terms of morality. However, a human might reply that this is a kind of cheat. It is easy not to lie to your spouse if you do not have one; coercive sexual acts are easy to avoid if there are no such things as sexual acts. The same is true with sexism. It is easy to avoid sexism if there is no such thing as sex. Still, it cannot be a moral failing of robots that they avoid many of our moral failings simply by not having the relevant requisite desires. There is some sentiment to the contrary in Western culture. A moral agent is seen as one who avoids temptation. But this is erroneous. The only reason we believe this is that we are all so tempted to do various bad things. Remove the temptations and then, as long as you still have agents, you still have morality. Indeed, perhaps the most moral being would be one who never thought about right and wrong, because it never occurred to it to do wrong. Whether or not one regards the robots as morally superior in light of their fewer temptations, the world of the robots is obviously a much better place than our world: their world is devoid of racism, sexism, rape, etc. True, some of these improvements are obtained cheaply (e.g. they have no sex), but this is part of why their world is a better place than ours. Finally, the robots will be autonomous and have desires, and hence they will almost certainly have conflicting desires. Therefore they will have temptations of their own to deal with. Hence they will have to make recognizably moral decisions, and they will also make mistakes. Still, they will behave much better than we do.

References

C.R. Chapman and D. Morrison, *Cosmic Catastrophes*, New York: Plenum Press, 1989.
D.C. Dennett, *Darwin's Dangerous Idea*, New York: Simon and Schuster, 1995.
E. Dietrich and V.G. Hardcastle, *Sisyphus's Boulder*, Philadelphia: John Benjamins, 2004.
T. Nagel, What is it like to be a bat, *Philos. Rev.*, 83, pp. 435–450, 1974.
S. Weinberg, *The First Three Minutes*, New York: Basic Books–Perseus, 1977.
E.O. Wilson, *The Future of Life*, New York: Alfred A. Knopf, 2002.