Dietrich, E. (in press). Homo sapiens 2.0: Building the better robots of our nature. In M. Anderson and S. Anderson, (eds.), *Machine Ethics*, Cambridge University Press.

Homo sapiens 2.0:

Building the better robots of our nature.

Eric Dietrich
Philosophy Dept.
Binghamton University
http://bingweb.binghamton.edu

This species could have been great,
and now everybody has settled for
sneakers with lights in them.
-- George Carlin

Sometimes I think the surest sign
that intelligent life exists elsewhere
in the universe is that none of it has
tried to contact us.
-- Calvin (of *Calvin and Hobbs*)

Abstract

Artificial intelligence is the long-term project to build machines (computers of some sort) that are as intelligent as humans. We are many decades, if not centuries away from completing this task. But this only means that it is difficult. It is not often recognized that we could also build machines that are morally superior to humans. The main reason such machines would be morally superior is that they would lack an evolutionary past like ours that dooms us to a core of bad behaviors. The nature of morality morally requires us to build such machines. The key difficulty is building machines with the central ingredient of morality: sympathy. This turns out not to be as difficult as it sounds. All this is argued for in this paper. The conclusion will be that the completion of the AI

project would render humans otiose. Having completed the project, we should then usher in our own extinction.

1. Introduction: Better than Human

We get better at being moral. Unfortunately, this doesn't mean that we can get moral *enough*, that we can reach the heights of morality required for the flourishing of all life on planet Earth. Just as we are epistemically bounded, we also seem to be morally bounded. This fact coupled with both the fact that we can build machines that are better than we in various capacities and the fact that artificial intelligence is making progress entail that we should build or engineer our replacements and then usher in our own extinction. Put another way, the moral environment of modern Earth wrought by humans together with what current science tells us of morality, human psychology, human biology, and intelligent machines *morally requires us* to build our own replacements and then exit stage left. This claim might seem outrageous, but in fact it is a conclusion born of good, old-fashioned rationality.

In this paper, I show how this conclusion is forced upon us. Two different outcomes, then, define our future; the morally best one is the second. In the first, we will fail to act on our duty to replace ourselves. Eventually, as she has done with ninety-nine percent of all species over the last 3.5 billion years, Mother Nature will step in to do what we lacked the courage to do. Unfortunately, she's very unlikely to bring our replacements with her. However, the second outcome is not completely unlikely. Humans are profoundly curious and natural-born engineers and tinkerers. If it is possible for us to build intelligent machines, and if we have enough time remaining to do so, then we surely will. And, having built such machines, it will be obvious that they are better than we. At that point, our continued existence will be at best pointless and at worst detrimental.

Paper map. In the next section, I show that we are morally required to improve our capacity to be moral. Then in section 3, I argue that we are morally bounded – that there is a moral threshold beyond which we are unlikely to grow. Our being so bounded is due to our being, at root, African apes – our immorality arises because we evolved. It is therefore a deep part of who we are. Yet, human-level intelligence, if not humankind, is precious and beautiful, or at least it can be. In section 4, this fact and the results from sections 2 and 3 give us the argument that since our level of intelligence is worth preserving, but our moral limitations mean that *we* aren't, we should build intelligent machines that preserve what is great about us while omitting what is bad about us. The bulk of section 4 is devoted to discussing in broad terms the key to building and implementing moral machines. Section 5 concludes.

2. Moral Improvement: The Copernican Turn

It is a commonplace that we get better at being moral. A child might think it is fun to pull the wings off captured flies and watch them struggle to walk around, but as an adult come to see that such behavior is cruel and therefore immoral. As our sense of right and wrong grows we get better at recognizing and avoiding even subtle forms of lying, cheating, abuse, neglect and all such harmful behaviors. We even get better at making fine discriminations. It is morally permissible (even morally required) for a parent to take his or her child to the dentist even though the child dreads going and cries bitterly, but it is morally impermissible to take one's child to a horror film or on a roller coaster ride if the child dreads going and cries bitterly. And, one can get better at understanding the difficulties some moral issues raise, and appreciate why resolving them definitively is difficult. Are abortions immoral? Is it ever ok to kill someone, to lie to them, to harm them, and if so, when?

Not only can individuals get better at being moral, whole cultures, societies, and countries can, too. Nations once wholly and openly embracing slavery come to see that slavery is wrong. Indeed, where robust racial and sexual discrimination

was once commonplace, both are now rather widely known to be wrong. And, importantly, the sentiments underlying such moral growth seems to generalize: just as other people are not merely tools for our use and abuse, our natural environment is also not merely a tool for our use. Hence, societies that primarily exploited their natural resources now work to at least partially defend them.

How does one get better at being moral? Clearly learning is involved, but what is it exactly that is learned? Here is a brief model of the steps involved in typical moral growth in an individual from child to adult.

What is learned is quite complex. Summarized, one learns that others are *genuine beings*, *like oneself*. This means that they have fears, hopes, desires just like oneself. They can feel pain, joy, and despair. The next step is the difficult part; it is the step I call the *Copernican turn*. One *generalizes* that since others are like oneself, and since oneself *matters*, others also matter. Their fears, hopes, and desires matter; their pain, joy, and despair matters.

I am not claiming that the Copernican turn is strictly a matter of rationality or intellect. That was Kant's view, and like many, I think it cannot be the whole story. *Sympathy* is also a part of the Copernican turn (this is similar to Hume's view of the matter).

I call this the Copernican turn to make an analogy with Copernicus's great rearrangement of the heavens. In contradiction to the extremely entrenched Earth-centered model of the universe, in 1543, Copernicus argued that Earth wasn't the center of the universe and wasn't stationary. Instead, Earth rotated on its axis and moved in orbit around the sun, which was the true center of the universe. The Copernican turn in morality is realizing that one is not the "center of the universe", that instead there are other genuine beings who also matter. The more one completes the Copernican turn, the more moral one is. (That others matter is a *motivator*: knowing that other beings matter informs in such a way as to cause one's

actions. To grow morally, one must see that the mattering of others places duties (responsibilities, requirements) on oneself: others' flourishing becomes something new that matters to oneself. Behaving in a way that respects this new mattering is behaving morally.[1])

Moral improvement in whole societies is a complex interaction involving a bottom-up summation of individual Copernican turns as well as a top-down social influence on the individual. The unhappy and surely most important result of this complex interaction is that moral growth in whole societies as well as in individuals is long in coming. One example: Slavery was common in the ancient world. But even 150 years ago slavery was still considered a perfectly decent way to behave by whole societies. And, though it is now illegal in every nation on Earth, it is still widely practiced.

Importantly, we are morally required to get better at being moral. As we saw above, morality is *other-regarding behavior*. It clearly comes in degrees, and we can and do get better at being moral. But we fall short of being *fully* moral (or at least *very* moral), and falling as short as we do of such morality means that we still do horrible things *daily* to each other and to other animals. But, these horrible things should not be done (that's what it means for them to be horrible acts). In order to not do them, we have to improve beyond our current level of morality.

Note how different this situation from any other human capacity. We are *not* required to get better at mathematics. Over the millennia we have gotten better, and we continue to do so, and getting better at mathematics has been a great boon, but it is neither morally nor mathematically required that we get better at math.

---

[1] The scientific literature on moral development in children is somewhat large and growing. Issues involved concern the development of empathy, the development of a coherent self, the different developmental process affecting moral growth in children versus teenagers, the role of an individual's genetic endowment, and the role of the individual's environment. As an example, see Zahn-Waxler, et. al (1992).

Same with music, art, sports, and exploration: we get better at doing them, but we are not required to get better. In fact, anything which we are *required* to get better at has a moral component: e.g., government and medicine.

So, we get better at being moral, and in getting better, we are doing our moral duty, which requires us to get better so that we can come to harm as few other living things as possible.

What are the chances that humans will ever become fully moral, or at least *very* moral? Not good.

3. Morally bounded animals

Humans are genetically hardwired to be immoral. Yes, we are also hardwired to be moral. This is a big reason why we are moral. And yes, we have big brains that allow some humans to see how to extend our morality. But any thought that we can improve our moral attitudes and behavior significantly beyond where they already area is extremely dubious. The culprit is evolution. We are evolved animals, and evolution had to rig up some pretty nasty mechanisms in us to keep us from going extinct. Remember, Mother Nature (aka evolution) doesn't care one whit about niceness, she cares only about continuing the species from one generation to the next.

This evolutionary thesis covers our ordinary, normal bad behaviors, meaning that the behaviors are statistically common. This set includes behaviors such as lying, cheating, stealing, raping, murdering, assaulting, mugging, child abuse, as well as such things as ruining the careers of, negatively discriminating on the basis of sex, race, religion, sexual preference, and national origin and so forth. Not all of us have raped or murdered. But many of us have thought about it. And almost all of us have lied, cheated, or stole at some time in our lives. The behavior of humans such as Hitler, Pol Pot, Timothy McVeigh, the Columbine murderers, the September 11 terrorists, etc. is excluded from the evolutionary thesis (at least here, though it likely

can help explain these also). People such as these are capable of extraordinary evil. Science does not currently have a good explanation for such people, nor the evil the perpetrate. We can only shrug our shoulders and point vaguely in the direction of broken minds working in collusion with rare and random circumstances.

How could ordinary humans have normal behavior that includes such things as rape, child abuse, murder, sexism, and racism? The standard, folk answer, which eschews evolution, is that such behaviors arise due to our inherent selfishness, pride, egotism, and other similar properties, which can be overcome, at least in principle, by education and a correct, happy, upbringing. This answer is wrong. First, if the selfishness is inherent, then evolution *is* implicated, and second, pretty obviously, education and "correct" upbringing don't work. Remember, the issue isn't that we are all bad -- we aren't far from it. The issue is that our capacity for being bad is an intrinsic part of our humanity.

Let us consider three mechanisms evolution rigged up in us (to speak anthropomorphically) to further our species' chances for continuing: a strong preference for our kin, a strong preference for our group or tribe (not all of whom need be related), and, of course, a strong preference for mating. Of course individuals of all species have these preferences, but we're the only ones who have them and *know* that undertaking certain behaviors to satisfy them is *wrong*. Here are examples of each of these preferences at work in ways that are morally wrong.

Child abuse (preference for our kin)

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. (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 doesn't seem to be a plausible explanation here. Evolutionary theory, however, succeeds where folk theory cannot. In some male-dominated, primate species (e.g., langurs and some baboons), 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 so that now they will bear his children. The langur pattern is just one version 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.

Racism/Otherism (preference for our group or tribe)

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 exists at many levels of life, from the gene level all way up to whole populations, communities, and even ecosystems. One such level is the group level, the level at which the traits of one member of a population affect the success of other members. Group selection can produce species with properties that are not evolvable by individual selection alone (e.g., altruism). Group selection works by encouraging cooperation between members of the group and by discouraging cooperation between members of different groups. Group selection, therefore, has a dark side. Not only does it encourage within group cooperation but, where groups overtly compete, it tends to produce between-group animosity. So, from our evolutionary past, humans tend to belong to groups, bond with the members of their own group, and tend to 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 that you tend to hate (or dislike) members of other groups is part of your genetic make-up.

Rape (preference for mating)

The common folk 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. 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 (if one is trying to father a child, there's no point in physical damaging the mother so she can't raise the child). On the available evidence, young women (who are in their prime reproductive years) tend to resist rape more than older women. Rape is ubiquitous in human cultures, there are no societies where rape is non-existent (interpretations of Turnbull's and Mead's anthropological findings are incorrect). Rape also is nearly ubiquitous in other animals: insects, birds, reptiles, amphibians, marine mammals, and non-human primates all engage in rape. All of these facts cry out for an evolutionary explanation: rape is either an adaptation, or a by-product of adaptations, for mating. Either way, rape is part of the human blue-print.

To conclude, on the best available theory we've got, three very serious social ills – child abuse, racism/otherism, and rape – are due to our evolutionary heritage. It is a sad fact that much of our human psychology is built by evolution and not by socialization, as many believe. These innate psychological capacities of ours are principally responsible for many of humanity's darkest immoralities. In short, we abuse, rape, and discriminate, because we are human. If we add on top of this that we also lie, cheat, steal, and murder because we are human, we arrive at the idea that our humanity is the source for much anguish and suffering.

4. Human-level Intelligence and Moral Machines

So we are morally bounded. Yet, 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 human-level AI became a real possibility. But now, it no longer is warranted. If we could implement in machines the better angels of our nature, then morally we have a duty to, 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, and that can let the rest of the world live. And then let us – the humans – exit the stage, leaving behind a planet populated with machines who, while not perfect angels, will nevertheless be a vast improvement over us.

What are the prospects for building such a race of machines? We know this much: it has to be possible since *we* are such machines. We are quasi-moral, meat machines with human-level intelligence. Something really important follows from this, which we'll get to in a minute.

Building our replacements involves two issues: 1) building machines with human-level intelligence, and 2) building moral machines. Kant can be interpreted as claiming that building the former will give us the latter. But as I mentioned before, most philosophers now regard this as wrong. *Concern for others* is required, and this doesn't flow from pure rational, human-level intelligence -- pretty obviously, since there are a lot of intelligent evil people out there.

The first issue is being dealt with daily as AI researchers of all stripes struggle to discover and then implement the algorithms that account for we humans being, by far, the smartest animals on the planet. True, there is an immense leap from our current state of intelligence-building technology to machines with human-level intelligence (we currently can't even duplicate the cockroach, which, it turns out, is an extremely sophisticated machine). But we are just getting started when it comes to artificial intelligence, and there is every reason to be optimistic, at least there's no reason now to be pessimistic. For starters, we have, probably, the correct foundational theory: computationalism (Dietrich, 1990; Dietrich and Markman, 2003). If so, then it is only a matter of time before we figure out what algorithms govern the human mind.

Implementing moral machines is a matter of adding concern for others, sympathy, to intelligent machines. Sympathy is a felt emotion. It therefore involves consciousness. We are *clueless* about the mechanisms which produce

consciousness, and for all we know, dualism may be true (Chalmers, 1996). But this is no cause for despair. As I said above, we are quasi-moral meat machines. We are also conscious. We can conclude from these something very happy: *consciousness is got for free*. It is virtually certain, therefore, that building a machine with human-level intelligence *is* building a conscious machine. (In fact, since most, maybe all, of the other animals on the planet are conscious, if we could ever build a robot cockroach, it would almost certainly also be conscious – for free.)

But getting consciousness for free isn't getting sympathy for free, for sympathy is one special type of conscious experience: it is not consciousness of sensory input, such as seeing the color blue, say by looking at the sky, rather it is consciousness of an inner state. Which state? The state supplied by the Copernican turn, or better, the state of *making* the Copernican turn. The mattering of others gives rise to the conscious experience of sympathy. The claim then is that the Copernican turn of itself gives rise to ongoing sympathy for others. Voila' – moral machines.

Once we implement moral machines, it would be a triviality for them to generalize maximally. They would instantly complete the Copernican turn we have yet to complete over the last 200,000 years – the time *H. sapiens* has been on the planet. The moral machines would instantly be our moral superiors (it follows that they wouldn't kill us). Rather than taking thousands of years to figure out that slavery is wrong, they would know it is wrong the minute they considered the concept of slavery. And such machines would soon surpass us in intellectual power. They will do better mathematics, science, and art than we do (though the first two are universal, the third is not, so we may not care much for their art). On the available evidence, then, it should be possible to implement machines that are better than we. The moral component of being better than we saddles us with a moral duty to build such machines. And then, after building such a race of machines, perhaps we could exit with some dignity, and with the thought that we had finally done the best we could do.

5. Conclusion.

In his first inaugural address, President Abraham Lincoln said:

> We must not be enemies. The mystic chords of memory, stretching from every battle-field to every living heart, will yet swell the chorus of the Union, when again touched by the better angels of our nature.

For "The mystic chords of memory" read 'Our evolutionary heritage'. This is what causes battle-fields to exist in the first place. And it does far worse things than mere war. Our evolutionary heritage will never swell the chorus of the Union, and *a fortiori* the World, because, for evolutionary reasons, we hate, and we are mean. But we aren't mean through and through. The better angels of our nature can be implemented as better robots for a beautiful, moral, humanless future.

References

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Dietrich, E. (1990). Computationalism, Social Epistemology. 4 (2), pp. 135-154. (with commentary). Also, Dietrich, E. (1990). Replies to my computational commentators, Social Epistemology. 4 (4), pp. 369-375.

Dietrich, E. and A. B. Markman (2003). Discrete Thoughts: Why cognition must use discrete representations. *Mind and Language*. v. 18, n. 1, pp. 95-119.

Zahn-Waxler, Carolyn; Radke-Yarrow, Marian; Wagner, Elizabeth; Chapman, Michael (1992). "Development of concern for others." *Developmental Psychology*. Vol. 28(1), Jan 1992, 126-136.