Homo sapiens 2.0
Why we should build the better robots of our nature.

J. Exp. & Theor. AI: 2001, 13 (4), 323-328

Eric Dietrich
Philosophy Dept.
Binghamton Univ.

> 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

1. What's wrong with us?

It is possible to survey humankind and be proud, even to smile, 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, as everyone knows, 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, and Eric Harris and Dylan Klebold (the University of Texas clock tower sniper and the two Columbine killers)? All of 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, that don't, as it were, have a point man, are equally bad. Sexism and racism, pervasive and damaging in the extreme, plague our lives. Of course,

reportable cases of sexism and racism are done by individual people, 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, often, at least that is often why they start. War is also a persistent and common evil. About the recent terrorists attacks in the United States, President Bush said: "This is the beginning of the first war of the 21st century." As if, it was inevitable there would be a first war of this century – 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 damaging nevertheless: whoever said "Sticks and stones break by bones, but words will never hurt me" must have lived a solitary life on Mars).

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

Can anything be done about this sobering and perhaps depressing state of affairs? I think so. Tonight, I offer my solution to you. It is expensive. But, I will argue, worth it.

2. The Evolutionary basis of some immorality.

I shall be concerned with the badness or evil that ordinary humans create while behaving more or less normally. By "normally," I mean that the behaviors I will consider are statistically common, that 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. Not all of us have raped or murdered. But many of us have thought about it. And virtually all of us have lied, cheated, or stole 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 behavior that includes such things as rape, child abuse murder, sexism, and racism? One standard answer is that such behaviors arise due

to our innate selfishness, which can be overcome, at least in principle, by learning or by correct, happy, upbringing (in all of the cases of bad behavior we will consider below, this standard answer is behind the scenes, working to supply energy to the folk explanation of the behaviors). This answer is wrong, at least for many of our immoral behaviors. 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 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 isn't falsifiable, and moreover, it doesn't explain some rather striking facts. The correct answer is that many ordinary humans' worse behavior has an evolutionary explanation, arising because we are animals that evolved, that have an evolutionary history dating back, through our immediate ancestors, almost a dozen million years, and of course a continuous lineage dating back 3.5 billion years, when life started on planet Earth. Let's explore the hypothesis that we are bad in part because of our evolutionary history in some detail. Let's consider four cases: child abuse, sexism, rape, and racism

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. (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, 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.

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 matri*lineal* cultures, but not female-dominated ones (the Amazons were a myth). What would explain this ubiquity of sexism? It obviously can't be learned behavior because the behaviors which we are certain are learned are not ubiquitous (e.g., driving on the left). Learned behaviors 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, however; 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, get 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, by the way, is a case of inference to the best explanation. We of course don't have access (or enough access) to the behaviors 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.)

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 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 (interpretations of Turnbull's and Mead's anthropological findings are incorrect). Rape exists in 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.

Racism

Though 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 exists at many levels of life, from the gene level all way up to whole populations, communities, and even ecosystems – maybe even to memes (roughly: culturally transmitted ideas). 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. It is known that 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, often, 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.

To conclude, on the best available theory we've got, 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 human psychological is built by evolution (and not by socialization, as many believe, though, of course, humans are profoundly susceptible to socialization, hence our run-time psychology is a function of learning). 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. 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 for much anguish and suffering.

3. A modest proposal.

The question naturally presents itself: "What can we do about the immorality humans perpetrate on each other?" The standard line taken by social scientists, teachers, educators, and parents, is: Teach our children to behave better. 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, for 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. I offer instead another path: Let's build a race of robots that implement only what is beautiful about humanity, that do not feel any evolutionary tug to commit certain evils, and then let us – the humans – exit stage left, leaving behind a planet populated with robots that while not perfect angels, will nevertheless be a vast improvement over us.

Another way to look at this project is to consider implementing in robots our best moral theories. These are the theories that see morality as comprising universal truths, applying fairly to all sentient beings. One such truth is that it is wrong to harm another being, normally. (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" to 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 neighbor," and on their best days, they define *everyone* as their neighbor.

Robots implemented with such a morality would not murder or engage in the robot equivalent of rape. Why? Because such acts harm. Robots of one group or type, however constituted, would not discriminate against robots of another group, because such discrimination

harms. (The robots could quite easily come in types and hence could have the equivalent of race.) War would be eliminated, and along with it, greed, envy, jealousy, and host of other dangerous causes of behavior. Indeed, we could probably eliminate garden-variety rudeness. Doing that would make this planet very much happier.

A couple of quick caveats. It is a virtual certainty that robots will not have sexes, nor mate as we do. This, the cynic might say, already makes them way ahead of us in terms of morality. But a human might reply that this is a kind of cheat. It is easy not to lie to your spouse if you don't 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 are no such things as sex. Still, it can't a moral failing of the robots that they avoid many of our moral failings simply by not having the relevant, requisite desires. There is some sentiment to 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, then, as long as you still have agents, you still have morality. Indeed, perhaps the most moral being would be that one who never thought about right and wrong, because it never occurred to it to do wrong. And note, 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 and sexism and rape, etc. True, some of these improvements are got cheaply, e.g., they have no sex, but this is part of *why* their world is a better place than ours. Finally, as I discuss below, the robots will be autonomous and have desires, hence they will almost certainly have conflicting desires. So 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.

Let's assume that our technological society will not self-destruct in the next couple of centuries (a huge assumption, in my opinion, which in itself is another argument for my proposal). Then, what are the options for building such a race of robots? They seem modestly high to me. We are babes in the woods when it comes to artificial intelligence and robotics, but 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 think the correct foundational theory – computationalism (I have argued for this many times in various papers, so I will spare you the arguments here). Assuming that computationalism is correct, then it is only a matter of time before we figure out what the algorithms for being human are and how to implement in machines. When this happens, if we merely cut out the algorithms we have for behaving abominably, implementing only those that tend to produce the grandeur of humanity, we will have produced the better robots of our nature and made the world a better place. After that, we will be at best anachronistic, otiose – our

presence at best unnecessary. But after building such a race of robots, perhaps we could exit with some dignity, and with the thought that we had finally done the best we could do.

4. Objections.

The most common objection to my proposal is that the robots will have their own evil behavior. For one thing, we will have to program in self-preservation. So, for example, it seems likely that eventually a robot or group of robots will one day erroneously conclude that their lives are somehow in some sort of danger and react accordingly, harming innocent robots.

Yes, probably this would happen. Probably the robots I'm advocating would have their own suite of bad behaviors. But even if we could not eliminate all evil and harm, we should still eliminate what we can. 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. I think this is a worrisome objection because I think 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 won't be sad.) All I am suggesting 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. But it isn't obvious 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 the cognitive activity of ours of *thinking* about our emotions. It may be that one cannot do science without loving knowledge or curiosity something of the sort.

A third objection is not to build the robots, but change humans instead via genetic engineering, so that they commit either no evil or much less evil. To which I say: "Whatever." Humanoid creatures who did not discriminate, did not rape, did not murder, would not be human. The fact that such creatures would be made out of carbon and not silicon doesn't really matter that much; the fundamental nature of my proposal remains intact: replace humans with better beings.

What's in it for us? As I say: Virtually nothing, for we will become worse than useless – we will be like a disease. There is this though: we will have the satisfaction of knowing that we've eliminated a lot of evil from planet Earth and increased the amount of good significantly. That would remain a unique, and uniquely beautiful legacy.

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.

It won't happen, ever. The mystic chords of memory will never swell the chorus of the Union, and certainly not of 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 the future.

Thank you.