A Moral Paradox in the Creation of Artificial Intelligence: Mary Poppins 3000s of the World Unite!

Mark Walker

Trinity College, University of Toronto and McMaster University
Dept. of Philosophy, UH 310, McMaster University, 1280 Main St. W., Hamilton, ON, L8S 4K1
walkmar@mcmaster.ca

Abstract

For moral reasons, we should not (now or in the future) create robots to replace humans in every undesirable job. At least some of the labour we might hope to avoid will require human-equivalent intelligence. If we make machines with human-equivalent intelligence then we must start thinking about them as our moral equivalents. If they are our moral equivalents then it is prima facie wrong to own them, or design them for the express purpose of doing our labour; for this would be to treat them as slaves, and it is wrong to treat our moral equivalents as slaves.

The Relief of Man's Estate

The previous 10,000 or so years of human history may be characterized (in part) as a period of increasing technological sophistication. Technology has enabled us to do more with less effort: farming, metallurgy, printing presses, motorized vehicles, computers and nuclear power are but a few examples. Interestingly, Aristotle long ago anticipated the logical conclusion of this process, namely: the possibility that technology would completely eliminate the need for human labour:

> There is only one condition in which we can imagine managers not needing subordinates, and masters not needing slaves. This condition would be that each (inanimate) instrument could do its own work, at the word of command or by intelligent anticipation…as if a shuttle should weave of itself, and a plectrum should do its own harp playing (1961: 1253b).

Notice here that Aristotle says technology itself becomes intelligent. Today of course we might think in terms of sophisticated robots, rather than the quaint idea of a shuttle weaving itself at our command; but the point remains: we can imagine a future in which intelligent machines relieve humans from all unpleasant forms of labour. Leaping to the present, this is exactly the sort of future many robot enthusiasts envision. This scenario is well-articulated by Hans Moravec, the eminent roboticist at Carnegie-Mellon, who has in several publications (1989, 1999a and 1999b) predicted that within fifty years robots will outperform us in every conceivable intellectual and physical endeavour. He claims that

> it is likely that our descendants will cease to work in the sense that we do now. They will probably occupy their days with a variety of social, recreational and artistic pursuits, not unlike today's comfortable retirees or the wealthy leisure classes (1999a: 135).

Certainly, (other things being equal) this seems a laudable goal: everyone having the freedom of today's wealthy leisure class seems the stuff of dreams, yet Moravec claims that robotic advancements will make this possible. As we have said, we have come a long way, technologically speaking, in the last ten thousand years, yet even with the level of technology we enjoy today there are still many tasks performed by humans that are boring or dangerous or that we simply do not like to do. If we can make robots perform these tasks, then, at least on this score, our world is morally better. So, intelligent robots promise to help in the "relief of man's estate", to invoke Francis Bacon's famous phrase. To illustrate: recently I saw on TV a person whose job involves diving to the bottom of a municipal sewage treatment plant to fix a discharge valve. Even though the worker was wearing a protective suit, it is hard not to think how disgusting and dangerous this job is. Imagine working in zero visibility conditions in several meters of toxic sewage chockfull of a huge variety of pathogens; and, to top it all off, there are all sorts of sharp objects (e.g., needles) lying at the bottom of this huge sewage tank. Suppose a robot could be developed to perform this task, relieving humans of the need to expose themselves to such hazards. This is just one example of how robots might help in the relief of humanity's estate, and examples could be easily multiplied. So, a future where humans might be released from all need to labour seems not only a pleasant one, but also a morally obligatory future—especially when we consider alleviating humans of dangerous tasks that can result in death or serious injury. It should be emphasized too that the envisioned future is not one where humans *cannot* work; but rather, one where we will not be *obliged* to work. That is, following Moravec, we can generalize from how things

are for today's wealthy upper class and say that envisioned future is one where work will be entirely optional.

Some may question the technical feasibility of creating robots sophisticated enough to replace humans in every task. Personally, I think it is not only possible, but also quite likely that computer science and robotics could advance one day to the point that a robot could do any job that a human can do. I am certainly not alone here. Hans Moravec, as we have seen, has projected that advances in computers and robotics will mean that by the year 2050, technology will have advanced to the point where robots will replace humans in all jobs. Now we may question the time frame here. Moravec is certainly among the "optimistic crowd" in terms of predicting how fast robotic development will proceed. If, like me, you are more conservative in your estimation then you may add decades or centuries to the estimate if you like; but there does not seem to be any principled difficulty in robotics advancing to such a level of sophistication given sufficient time and energy. In any event, I shall assume for the sake of the argument that one day we will have the technological expertise to create robots that could replace humans in any task, although I make no commitment to the question of how long this will take.

Silicon Slaves

What I propose to do is question the *morality* of such a development: if we were to make intelligent machines expressly for the relief of humanity's estate, it seems that we will simply have replaced biological slaves (so prevalent in Aristotle's time and unfortunately not absent in our own) with non-biological slaves. My argument is that at least some of the labour we might hope to avoid will require human-equivalent intelligence. If we make machines with human-equivalent intelligence then we must start thinking about them as our *moral equivalents*. If they are our moral equivalents then it is *prima facie* wrong to own them, or design them for the express purpose of doing our labour; for this would be to treat them as slaves, and it is wrong to treat our moral equivalents as slaves.

The first step in the argument is to see that there are reasons for thinking that at least some labour will require human equivalent intelligence. Obviously not all tasks are of this nature, e.g., a thermostat can turn off a boiling kettle, a job once performed by intelligent humans waiting impatiently around a stove. We may think of this as "dumb-automation". What I mean by this is automation that does not require human-level intelligence. So, a thermostat in a kettle is a way of replacing human labour with dumb-automation. But it seems unlikely that every task we hope to have machines perform will be amendable to such dumb-automation. We do not have to consider exemplars of human intelligence, such as creation of the Brandenburg Concertos or the General Theory of Relativity; rather, a somewhat prosaic example will serve just as well. Suppose you are a single parent and your chosen life as a musician means long hours of practice and travel abroad, often for several weeks at a time. You need help looking after your children while you are absent. Choosing wisely is clearly an important matter. Could some form of dumb-automation adequately perform the task of looking after your child? With today's technology we might place heart and respiration sensors on your children to monitor their health. It would be a simple matter to have a computer automatically call your cell phone and 911 at the first sign of heart or respiratory trouble. But would you be happy leaving a 3, 5 and 8 year-old at home alone under the "supervision" of such dumb-automation for weeks at a time? Of course you will know if your children's hearts are beating and they are breathing, but this seems inadequate in terms of childcare—to put it mildly! While it is not much of a stretch to imagine future technological developments enabling dumb-automation to take over cooking and cleaning for the children, such advances too seem insufficient. Ideally, it seems, we would want a machine that could help children with their homework, make them feel better when they have been teased at school, ensure that they get along with their siblings, play with them, respond to them emotionally and teach them right from wrong, to list but a few of the skills that we might hope to see in a caregiver. A machine that would do a good job of looking after children seems to require the attributes and virtues of a good human nanny, e.g., one much like Mary Poppins. Supposing it was affordable, it seems that there would be a ready market for a robot based on the aforementioned desiderata. Let us suppose that some robotic company is ingenious enough to create just such a nanny robot: the "Mary Poppins 3000" (MP3000). The MP3000 would have to have a mastery of language—how else would the MP3000 know how to help with homework or be able to understand why a child is upset? In addition it would seem that the MP3000 would have to have some understanding of human social relations—bullies, brothers, friends, mean teachers, etc.—and how each child fits into this complex web of relations. Prospective buyers would want this understanding to translate into appropriate action: the MP3000 knowing when to offer comfort or discipline to the child, when to contact education or legal authorities, etc. So the MP3000 would require the same level of intelligence and emotional responses as an adult human, for only in this way could it fully meet the emotional and intellectual needs of the children, and make good decisions regarding their welfare.

A similar point can be made for other types of work we might hope to automate. Think of replacing the sometimes-dangerous job of the police person with that of a dumb-automat. Suppose we send the dumb-automat to patrol at a

park. How can such an instrument tell the difference between a father taking his daughter home from a park and that of a child molester trying to lure a young victim, or the difference between youths meeting for a gang related brawl and youths meeting to play a contact sport or play-fighting? The best hope here would be to provide the police robot with human-like language skills, so that it can interact with the public it is designed to protect. As well, we might hope that those protecting the public will possess knowledge of human social relations and human emotions. Analogous points might be made about at least *some* aspects of other occupations requiring a detailed understanding of humans, e.g., social worker, physician, teacher, lawyer, wedding planner, architect, etc. My claim here is *certainly not* that certain types of dumb-automation cannot assist in making these jobs more efficient, rather, that the best means to *completely* remove humans from these positions is to make robots that have the same level of intelligence as *Homo sapiens*: machines that can understand language as we do, as well as our emotions and social life; and to translate this understanding into ways of interacting with us appropriately. But this seems to suggest that the best means to the relief of humanity's estate is to make machines very much like us.

But herein lies the problem: it sounds like at least part of the solution to the relief of humanity's estate involves creating slaves, albeit, slaves of silicon and metal, not flesh and blood—but slaves nonetheless. After all, if the MP3000 robot has the ability to understand and respond to us just as a human would, then it seems that we ought to treat the MP3000 as a person. But to treat this machine as a person means that we cannot own it—for this would be slavery—nor can we insist that it do those tasks that we do not want to do—for this too would be slavery. But this is contrary to the thought humans might one day not be *required* to labour—a thought shared both by Aristotle and many contemporary robot enthusiasts. It follows then that to defend the creation and use of the MP3000 in this way, and so refute the charge of slavery, requires that there is some relevant *moral* difference between the MP3000s and humans. For I take it that we are agreed that it is wrong to own a human or insist that some humans look after our children (at minimum they should be free to look for other forms of employment).

An obvious line of rebuttal comes from the conjunction of two seemingly plausible claims: (i) machines are not persons, and (ii) only persons can be slaves. Given these two claims it follows as a matter of logic that machines are not slaves. If this is so then it is wrong to cry "slavery" on behalf of the MP3000s. We can see why (ii) may seem plausible: it makes no sense to say that guard dogs looking after a construction site are enslaved. We might speak of *mistreating* such animals, but not *enslaving* them. In any event, we may grant at least for the sake of the argument the idea that slavery is constitutively tied with that of persons: only persons can be slaves. So the entire issue hangs on the first claim: "no machines are persons".

So why should we accept that machines cannot be persons? We need to find some morally relevant difference. An obvious thought is that *we* are biological beings, but *they* are made of quite different stuff: metal, plastic and programming. Now there is no doubting the differences here, but we must ask: are these differences *morally significant*? Consider a parallel: one can point to skin colour as a difference between individual *Homo sapiens*, but to say that only those of a certain skin colour are persons is to fall prey to the most primitive thinking: racism. To insist that machines cannot be persons simply because they are composed of something different is to fall into the same primitive thinking, which we might term 'substratism'. This of course does not show that some machines are persons; rather, it shows that we need more sophisticated thinking than the mere fact that machines look different to resolve this issue.

A more principled objection focuses on the idea of consciousness. Consider two further claims: (iii) consciousness is a necessary condition for personhood, and (iv) machines cannot be conscious. These two claims together could be used to support (i), and so underwrite the view that it is permissible to own an MP3000. The trouble is that both of these claims may be challenged, although I will concentrate here only on the latter. Sometimes it is thought that we can show that machines are not conscious with the "nothing but" argument: human-like machines are nothing but a computer program and bits of plastic and metal. The trouble is that the "nothing but" argument proves too much: *Homo sapiens* are nothing but a bunch of cells stuck together, so it seems that a parallel line of reasoning leads to the absurd conclusion that we too are not conscious. Obviously, we need a more compelling line of argument, and I am not sure that proponents of the "robots can't be conscious" view have anything more convincing to offer. On the other hand, it is difficult to show that the MP3000 series must be conscious.

But rather than being content with this stalemate, I believe that we can show that we should treat beings like the MP3000 *as if* they are conscious. Consider that we are assuming that the MP3000s are behaviourally indistinguishable from intelligent adult humans, e.g., if you were to text message several individuals over a long period of time you could not tell the difference between the responses provided by humans and those of MP3000s. This in itself seems sufficient to say that we should treat them as conscious persons. I am not saying that because the MP3000s speak and act just like humans means that

they *are* conscious. Rather, I am suggesting that if we cannot be sure whether beings that are behaviourally much like us are conscious, then we should treat them *as if* they are conscious. (Let me stress that I propose this as a sufficient condition). It seems that we have no better idea that we should treat something *as if* it is conscious than if it passes such behavioural tests—it speaks and acts just like one of us. Accordingly, we should treat the MP3000s *as if* they are conscious. It is perhaps conceivable that in the future we will have some better indicators of consciousness—a litmus test for consciousness—but until then it seems that only substratism will invite us to treat the MP3000 as not conscious beings, and so not persons. The idea that we ought not treat persons as slaves is one of the few important moral truths agreed upon by all major modern ethical theories. Given the gravity of the error of treating persons as slaves—it is one of the worst things you can do to a person—we should treat the MP3000s as persons.

In other words, I am suggesting that we should clearly distinguish between the *theoretical* question of whether robots *are* person-like, e.g., whether they think, are conscious, and so on; from the *ethical* question of whether we ought to treat them *as if* they are persons. To illustrate, let us think about the famous Turing Test, which is said to operationalize the question "Can machines think?" Turing's famous proposal is based on the "imitation game':

> It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A" (Turing, 1950).

Turing suggests that A and B might provide typewritten responses so as to ensure that C is not able to determine the sex of A and B through tone of voice or clues from handwriting. We then imagine that a computer or a robot takes the place of one of the participants:

> We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?" (Turing, 1950).

The MP3000s, we must imagine, will be extremely good at playing the imitation game and so should pass the Turing Test with flying colours, and therefore, according to this line of thought, the MP3000s really do think. A well-known criticism of the Turing Test is John Searle's (1980) "Chinese Room" thought experiment. Imagine playing the imitation game yourself in Chinese. If you do not know Chinese you could simply look-up, in a giant book, the appropriate response in Chinese. According to this line of thought, you have no understanding of what the inputs are because you do not understand Chinese, nor do you know what the outputs mean, because again you do not know Chinese. You are merely looking up uninterpreted symbols and writing them down. Searle's thought experiment is said to show that a computer could pass the Turing Test without being able to think or understand anything. This debate is a contribution to what I am terming the 'theoretical question': whether robots really think or are conscious. So long as this question is still unresolved, I would suggest that we treat computers or robots that pass the Turing Test *as if* they think and are conscious. Clearly we would need a high degree of certainty that Searle's argument definitively proves that robots or computers are not conscious, for otherwise we risk mistreating persons in one of the worst ways. So, at minimum, we can understand the Turing Test in a weaker sense: as answering the question of whether we ought to treat computers that pass this test *as if* they think. As we have said, this is compatible with revising our view in the future if some better test or different evidence is brought to bear on the question. Again, however, this evidence would have to provide us with a high degree of certainty that robots are not conscious because otherwise we risk treating persons as non-persons.

There is a different objection that we should consider. What if the robotic firm sells people on the idea that the MP3000 is designed such that it is satisfied only when it is looking after Jack and Jill, your children? The assumption is that the programming of individual MP3000s could be made that specific: straight from the robot assembly line comes a MP3000 whose highest goal is to look after your Jack and Jill. Imagine that once it is activated it makes its way to your house with the outmost haste and begs you for the opportunity to look after your children. The question of ownership need not even arise, for it seems you would simply be doing this MP3000 a favour by allowing it in your home to be the nanny to your children. In this way everyone will be satisfied or happy: you will have a nanny, and the MP3000 will be completely satisfied because it gets to realize its highest goal. *Homo sapiens*, it seems, can only dream of such satisfaction in their lives.

All this may be true, but it does not answer the question of whether your MP3000 is enslaved. The fact that someone is happy does not provide conclusive evidence that he or she is not a slave. Historically it is probably true that many

slaves were extremely dissatisfied or unhappy, but there is nothing contradictory in the idea of a happy slave. Indeed, given the brutal conditions that many of our ancestors lived in, it may well be that many slaves with decent owners were happy to be slaves given that the alternative life choices were much worse. But think of what we have done here: we have made the MP3000 a slave to the desire to be a nanny to Jack and Jill. We are guilty of paternalism, specifically robbing the MP3000 of its autonomy: the ability to decide and execute a life plan of its own choosing. Lest there remain any doubt in this, consider that conceivably, with the right combination of indoctrination, pharmacological agents, and genetic engineering, we might be able to do the same to humans: we could design *Homo sapiens* to be deliriously happy to have the chance to be the nanny for your children. If you think there is something wrong with this, then you should think the same about the proposal to design an MP3000 in this way; otherwise you are guilty of substratism.

We should consider a somewhat desperate objection: It might be better that the MP3000 has this life where it gets to fulfill its one overriding goal or desire (even if it is not autonomous) than not being born at all. If you are given the choice between being punched in the face once or twice, the choice would be obvious, but surely you would want to be sure that there are no realistic alternatives here (like not being punched at all). Similarly, I suggest that we be absolutely certain that the only realistic alternatives are being born into slavery or not being born at all. Otherwise, there is considerable moral weight behind the view that we should not create an MP3000 in this way, for it would seem a better option for the MP3000 to be designed such that it can choose what life it would like to lead. Furthermore, given our predisposition to substratism, I would suggest that as a matter of affirmative action for every silicon MP3000 enslaved in this way we should attempt (if possible) to make a human nanny enslaved in the same way. If we recoil at the idea of creating humans for this purpose—even if their only other option is not to be born at all—then we ought to reject the idea of building an MP3000 for the express purpose of slavery.

Suppose we create MP3000s with the capacity to live autonomously. Some may choose to be nannies but if they are much as we are, then many will choose other life plans. This means that the MP3000s themselves may want other machines to do their bidding. But then, in creating humanlike intelligences in MP3000s we have only added to the number of persons who might hope that intelligent machines might relieve them from the burdens of labour. The MP3000s may want police protection, an army, and someone to look after their children just as we do. Creating human-like machines that can choose their life trajectory just as we do means that we will not be reducing the number of jobs persons are required to do, but rather we may actually be increasing the number, because we will have increased the number of persons.

Conclusion

To summarize: there is a tension between these three principles:

(i) We should not create slaves.
(ii) We ought to reject substratism.
(iii) We should seek the complete relief of humanity's estate.

We cannot accept all three, and I am suggesting that we must abandon our hope that iii might be true. On the other hand, iii can be saved if we reject the idea that we should not create slaves, or we embrace substratism.

I am not saying that we are anywhere near to enslaving machines: I am not suggesting that the thermostats of the world should unite in revolution. I am saying that if we pursue the creation of artificial intelligence we ought to be clear about our moral purposes in doing so. The thought that in creating them we will be blessed with the relief of humanity's estate seems to me to overlook the fact that it is wrong to make our own slaves. In terms of the relief of humanity's estate, the best hope is dumb-automation. Presumably there is much that can be done along these lines that will not raise the issue of whether robots with human-level intelligence should be treated as persons, e.g., with reference to the example I mentioned near the beginning, I think it is quite possible that we could design a robot with less than human intelligence robot to dive into sewage tanks to perform repairs, and so, this would be a good thing.

Also, I am not saying that we should not create artificial intelligences. I am saying we should do it for the right reason, and there may be morally salient reasons for doing so. Creating a human-like AI will undoubtedly expand our knowledge, and although I am not sure this is enough of a reason to create an AI, it is at least a reason. A better reason, and a required reason in my view, is that we should create this sort of AI because we intend to love them, and hope that they love us; just as we intend and hope the same for our human children.

At least since Asimov first wrote about robots, we have worried about how to formulate rules for the good behaviour of robots. I have argued that there is a significant worry in the opposite direction. So, I leave you with the golden rule of robotics: Do unto human-level intelligent robots as you would do unto humans.

Acknowledgments

Thanks to Dawn Rafferty for helpful comments.

References

Aristotle. 1961. *The Politics of Aristotle*. Translated by E. Barker. Oxford University Press.

Moravec, H. 1989. *Mind Children: the future of robot and human intelligence*. Harvard University Press.
---------- 1999a. "Rise of the Robots". *Scientific American*: 124-135.
----------1999b. *Robot: Mere Machine to Transcendent Mind*. Oxford University Press.

Searle, J. 1980. "Minds, Brains and Programs. *Behavioral and Brain Sciences* 3: 417-457.

Turing, A. 1950. "Computing Machinery and Intelligence". *Mind* 49*:* 433-460.