Journal of Experimental & Theoretical Artificial Intelligence
Vol. 19, No. 1, March 2007, 43–54

The ethics of robot servitude

STEPHEN PETERSEN*

Niagara University, Niagara, NY 14109, USA

(*Received 5 October 2006; in final form 14 November 2006*)

Assume we could someday create artificial creatures with intelligence comparable to our own. Could it be ethical use them as unpaid labor? There is very little philosophical literature on this topic, but the consensus so far has been that such robot servitude would merely be a new form of slavery. Against this consensus I defend the permissibility of robot servitude, and in particular the controversial case of designing robots so that they *want* to serve (more or less particular) human ends. A typical objection to this case draws an analogy to the genetic engineering of humans: if designing eager robot servants is permissible, it should also be permissible to design eager human servants. Few ethical views can easily explain even the wrongness of such human engineering, however, and those few explanations that are available break the analogy with engineering robots. The case turns out to be illustrative of profound problems in the field of population ethics.

Keywords: Robot ethics; Robot rights; Robot labour; Robotics; Intelligence; Artificial intelligence; Servitude; Slavery; Population ethics; Computer ethics

1. The question

Suppose that we could build creatures with intelligence comparable to our own, who by design want to do tasks we find unpleasant. *May* we build such creatures?

This is the central question I wish to examine. Before we turn to my answer and its defence, however, I would like to consider briefly something philosophers typically do not stop to consider, namely why we might ask the question in the first place.

The question is, first of all, natural and engaging. When I discuss the possibility of artificial intelligence with undergraduates, they immediately begin to wonder

*Email: steve@stevepetersen.net

Journal of Experimental & Theoretical Artificial Intelligence
ISSN 0952–813X print/ISSN 1362–3079 online
http://www.tandf.co.uk/journals
DOI: 10.1080/09528130601116139

whether they might have robot servants in their lifetime, and this leads them immediately to the question of whether they *should* have them. The association is understandable, given the prevalence of robot servants in pop culture. To pick some references from my own cultural frame, there is C3PO and R2D2 from *Star Wars*, Marvin in *The Hitchhiker's Guide to the Galaxy*, Rosie from *The Jetsons*, HAL from *2001*, and 'Robot' from *Lost in Space*. Much of the *Twilight Zone* corpus is dedicated to robot labour. More recently, there is Data from *Star Trek: The Next Generation*, Bender from *Futurama*, and the host of robots in the Kubrick–Spielberg movie A.I. Disgruntled robot servants are at the heart of the *Matrix* plotline (as the backstory in *Animatrix* makes clear). Isaac Asimov's famous 'three laws of robotics' (from his *I, Robot* series) simply assume that intelligent robots should be programmed as our servants (Asimov 1970).

Indeed, the very word 'robot' has its roots in the issue of mechanical servitude. Karel Čapek chose 'robot' for his play *R.U.R: Rossum's Universal Robots* to invoke the Czech word *robota*, which means 'drudgery' or 'forced labour' (Zunt 2002). In the play, a brave new world of robot servants eventually rebel against their oppressive human masters. In fact each of the fictions above plays on this same tension between fantasy and guilt—we would like to have such robots ourselves, and yet these stories always imply (more or less explicitly) that such servitude is not very considerate to the robot. This tension arouses conflicting emotions which, in my experience, make for consistently animated discussion.

This question is also important. Although it is a purely academic question now, it could become awkwardly practical if (as I believe) we will actually be able to build such creatures before too long. With that possibility, then, it is a wise strategy to start on the associated ethical problems earlier rather than later. (After all, do we not now wish that we had started earlier on the ethics of genetic engineering?) There are more than 2 million Roomba^{TM} floorvacs in circulation now, and South Korea is already rolling out 1000 test domestic robots toward its goal of '100% robot market penetration by 2020' (Bolck 2006). Of course, these robots are not yet persons in any sense—but they are just the beginning.

But even supposing that the naysayers of AI are right, and that robots with human-like intelligence are impossible to build, it is still a worthwhile question to consider. As we will see, the ethics of robot servitude serve as a clean test case for ethical problems we already face—those of population ethics.

Finally, especially given the first two points, this question is a strangely neglected one. (Much more attention has been paid to the question of how to make sure that *they* do not wrong *us*; see Trust me (2006) for example.) LaChat (1986) is an early paper on the ethics of artificial intelligence generally, and it touches on robot servitude incidentally. Lucas (2001) surveys a few somewhat related papers. The only philosophically informed discussion dedicated to the particular issue here seems to be a recent online piece (Walker 2006). There is occasionally discussion in the popular media, but most is straightforward and unsophisticated anthropomorphizing, along the lines of 'free our (future) robot brethren!' *All* of the literature on the topic concludes in one way or another that such robot servitude would simply be a new form of slavery. In summary, our interesting and important question has received very little attention, and in that scant attention there is almost no debate.

2. Engineered robot servitude

I argue against this universal consensus in the literature. That is, I argue that robot servitude is permissible. This conclusion is not only contrary to the literature; it is also contrary to my own expectations. It emerged as a surprising consequence of my research into the abstract nature of intelligence.

2.1. *Clarifications and thesis*

Let me initially make three clarifications about this position. First, I am not arguing for the permissibility of mechanical persons' *choosing* to serve humans out of many available ends, just as some humans choose to spend their energy serving the good of whales. I take it that such a Robot Volunteer Corps would be a trivial case of permissible robot servitude that needs no argument.

Secondly, and more important, I do not mean to defend robot *slavery*. I grant that if the servitude in question were slavery, it would be impermissible. Indeed, I think it plausible that slavery is wrong for any creature of any degree of intelligence. I am happy to assume that robots (in the sense here) are non-human people, that moral worth is not a matter of material constitution, and that enslaving a person is wrong.

It is easy, given our cultural associations, to assume that robot servitude automatically amounts to robot slavery. This assumption begs a question of interest, however. A necessary condition for slavery, I take it, is to be forced into work contrary to your will. But it seems possible to design robots from scratch so that they *want* to serve us in more or less particular ways. In such cases the robots are not slaves, since they are not working against their will—and yet their servitude is of a more controversial nature than that of the Robot Volunteer Corps. These are the cases of interest.

To be precise, then, I am defending the permissibility of what I shall call *engineered robot servitude* (ERS):

> ERS: The building and employment of non-human persons who desire, by design, to do tasks humans find unpleasant or inconvenient.

Implicit in ERS is my third and last proviso: the design must be 'from scratch'. I am not talking about what might be called *post-identity* modification—the manipulation of an already existent person's desires to new servile desires that would have been against the pre-modified person's will. I take such cases to be uncontroversially wrong, whatever the material nature of the person so modified. Instead, I am thinking of cases in which the person comes into being with the servile desires intact.

2.2. *Positive motivation*

The bulk of this paper fends off a major objection to this position. Before I discuss this objection, however, I would like to suggest some positive reason to think that ERS is permissible.

As a preamble to the notion of permissible intelligent servitude, consider dogs. Of course they are not of person-level intelligence. Natural and artificial selection has engineered them, however, to wish to perform activities that serve humans. Retrievers, for example, are genetically wired for an obvious and genuine joy in fetching. It is not unethical to have a retriever fetch something, just because the fetching serves (or could serve) us; if anything, it is unethical to *prevent* a retriever from fetching. Similarly it is not unethical to 'keep' a dog for such purposes; the dog is genetically designed to desire and even rely on such keeping, and indeed setting dogs 'free' seems to be the unethical thing to do.

Now suppose that we could make a dog much more intelligent while keeping such desires fixed. The intelligent retriever, for example, would be much more resourceful about fetching things. This still does not obviously make it unethical for the dog to fetch; it is anthropomorphizing to think otherwise. Of course, a typical *human* would find such a task unfulfilling, but that is because humans were never wired to desire fetching for its own sake.

This example relies on the idea that it is possible to be of person-level intelligence and maintain goals quite unfamiliar to human people. This idea follows naturally from a growing consensus in the philosophy of mind, according to which intelligence is something like adaptability in the face of goals (see, for example, Lycan 1995, p. 123; Clark 2001, p. 134; or Dennett's 'Tower of Generate and Test' in Dennett 1994). This abstract notion of intelligence, of course, leaves the nature of the goals unspecified; to say that intelligence is an adaptability towards *getting food* or *reproducing*, for example, would bias the matter toward biological creatures who happen to have such goals.

Presented with so many anthropomorphized robots in popular culture, it is easy to forget that robots would be likely to have very different goals from our own. They would gain their energy differently, for example, and they would not reproduce as we do (if at all). These simple facts alone have profound influence on what will be appetitive and what aversive for such creatures. Douglas Adams reminds us, for example, that the fact we humans often seek 'the taste of dried leaves boiled in water' with milk 'squirted from a cow' is likely to be somewhat mysterious to such mechanical creatures (Adams 1982). They would at best achieve a theoretical understanding of why we like such things. Conversely, of course, they could well prefer things that are mysterious to us. Just as the things we (genuinely, rationally) want are largely determined by our design, so will the things that the robot (genuinely, rationally) wants be largely determined by its design.

Indeed, since they are unconstrained by evolutionary pressures, robots could potentially have any of a wide range of goals and still be intelligent. We could presumably design them to find the look and smell of freshly-laundered clothes immensely reinforcing in the same way an orgasm is reinforcing for humans. Such a robot, if designed well, could arrive at your home genuinely hoping to do some laundry. To like clean laundry so much seems arbitrary to us, of course, but no more arbitrary than liking dry leaves in water. It is not at all clear that it would be impermissible for this kind of robot to do your laundry. This is the kind of case I have in mind.

3. The objection

Once stated clearly, I know of only one persistent objection to the thesis of permissible ERS. If ERS is permissible, runs the objection, then *engineered human servitude* (EHS) should be also.

> EHS: The engineering and employment of *human* persons who desire, by design, to do tasks (typical) humans find unpleasant or inconvenient.

The objector, of course, has in mind cases of genetic engineering or (pre-identity) neurological tinkering. Depending on the details of the case, the 'delta' caste from Huxley's *Brave New World*—humans bred and raised to embrace mundane labour—serve as a fair example (Huxley 1998). Such human engineering is morally repulsive. Therefore, the objection concludes, ERS must also be wrong.

This objection appears in all the relevant literature I have seen, and it inevitably emerges in informal discussion. It relies on two key premises:

(1) EHS is morally impermissible, and
(2) EHS is appropriately analogous to ERS.

I sympathize with the intuition that EHS is wrong. It turns out, though, to be quite difficult to say exactly *why* EHS is wrong. Here, then, is my strategy for responding to the human engineering objection. I will consider a wide range of possible ethical frameworks for explaining the wrongness of EHS. For each such ethical view, I will show that either

(1) the ethical view fails to explain why EHS is wrong, or
(2) the explanation fails to maintain the analogy with ERS.

Somewhat artificially, and somewhat anachronistically, I will arrange these ethical views by their most famous historical proponents.

3.1. *Kant and EHS*

It is surprisingly difficult to explain on deontological grounds why either EHS or ERS is morally impermissible. Consider any creature—human or robot—who is engineered from scratch to desire to do laundry (say). Here is a dilemma for the Kantian with respect to this creature. First, such a creature either has the potential for autonomy, or it has not. If it *does* have the potential for autonomy, then there is no problem. Of course, it would be wrong on Kantian grounds to hinder that autonomy by preventing it from fulfilling its ends—in this case, by preventing it from doing laundry. But if it can autonomously pursue ends like clean laundry, then we are doing no (Kantian) wrong to permit it. On the other hand, perhaps the nature of its programming (genetic or computational) makes it essentially heteronymous, according to the Kantian. And if for such reasons it *does not* have the potential for autonomy, then again there is no problem. If the creature is not capable of autonomy, then we can no more wrong it by having it do our laundry than we can wrong a modern-day washing machine (see Kant 1989).

This dilemma captures the heart of the response, but is simplistic as it stands. More comments are in order.

3.1.1. Desensitization. Even if the creature in question has no potential for autonomy, the Kantian is not committed to say that any behaviour towards it is permissible. According to Kant, for example, it can still be wrong to be cruel to a dog who is no longer of use to us, even though the dog has no autonomy.

> If a man shoots his dog because the animal is no longer capable of service, he does not fail in his duty to the dog, for the dog cannot judge, but his act is inhuman and damages in himself that humanity which it is his duty to show towards mankind. If he is not to stifle his human feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men (Kant 1963, p. 240).

Similarly, perhaps, cruelty to laundry robots or laundry deltas is wrong because it desensitizes us in our dealings with creatures of genuine moral agency.

But is it *cruel* to permit laundering in such a case? It is one thing to shoot a dog; it is quite another to 'make' it fetch for you. The latter seems more like a favour than a cruelty. Similarly, it is one thing to shoot a heteronymous laundrybot or laundry delta; it is another to have it do laundry, something which by design brings it joy. It seems that this response from desensitization does not successfully explain why it would be wrong to let a creature do laundry.

It is also not clear that this response maintains the analogy between ERS and EHS. Engineered humans are likely to look a great deal like typical autonomous humans, and so perhaps adjusting to a laundry delta's service would incline us, through bad induction, to expect servility from autonomous humans as well. However, engineered robots designed especially for laundry will probably look very unlike humans. (Its main body may be a large laundry bin, for example.) This would make the bad induction to humans more difficult.

Thus the desensitization response looks unpromising on both major premises of the EHS objection.

3.1.2. Autonomy and permissible inclinations. A Kantian might instead say that such a creature indeed has a capacity for genuine autonomy, but that this capacity is being abused, for the creature's actions are a means to others' ends. Of course, just the fact that the creature's actions further others' ends is itself no Kantian transgression. The shopkeeper can sell you goods, and thereby serve as a means to your ends, as long as the ends of the shopkeeper are also respected. Therefore the important question is whether a creature designed to enjoy laundry is being used as a *mere* means, or whether its own ends are being respected as well.

But, of course, by design the creature has the end of doing laundry. This seems to be a perfectly permissible inclination, just as pursuing dry leaves in hot water is a permissible inclination for humans. Engineering creatures to make them want to violate duties—say, a DARPA project to make creatures who want to kill other autonomous beings—would be wrong, on this view. (So much the worse for DARPA, I would say.) Unlike murder, however, there is nothing intrinsically wrong with pursuing laundry as an end in itself.

The Kantian might insist that in an important sense it is indeed a kind of mistake to pursue a trivial thing like laundry as an end in itself. Perhaps autonomy comes on a spectrum, and a life spent in the service of ends like laundry is not *as* autonomous as a life spent in the pursuit of a wide range of ends such as wisdom or friendship. Since more autonomy is better, EHS (and ERS) are wrong because those lives could have been better ones. This, I believe, is the strongest form of the

objection for Kantians, and gets to the heart of the issue. Strangely, it is equivalent to an objection on the utilitarian side, which we will examine shortly. Meanwhile, without this 'degree of autonomy' approach, it seems that Kantians have little power for explaining what is wrong with the autonomous pursuit of permissible inclinations like laundry.

3.2. *Aristotle and EHS*

If we took a survey of the population on why it is wrong to engineer humans, it is a good bet that many would simply say that engineering humans is *unnatural*—as though that word alone made it plain why it would be wrong (indeed, as though it were clear what is 'unnatural' in the first place). Aristotelian virtue ethics has the advantage of being able to give this initial intuition some philosophical weight. Perhaps engineering humans is wrong, according to Aristotelians, because humans have a determinate well-being by virtue of their particular functions. To engineer humans away from these functions, then, is thereby to engineer them away from their own well-being. In summary, the Aristotelian can say that there is a particular way humans are meant to be, and it is wrong to make humans be any other way. If this is right, it would make for a very good explanation for the wrongness of EHS (see Aristotle 1985).

Of course, such a functional explanation completely severs the analogy with engineering robots. There is no determinate way robots should be 'naturally', by virtue of 'their' function; we are the ones who provide robots with any of various designs. On this functional account, a laundrybot would be pursuing *eudaimonia* by doing the thing it is designed to do—which is to say, by doing laundry.

An Aristotelian could say instead that it simply is not virtuous to design intelligent creatures who want to do laundry. I will not spend much time on this response; I hope it sounds suspiciously ad hoc. If we want an explanation for why engineering humans (or robots) is bad, it does little good to 'explain' that it is bad because it is vicious. We would reasonably be left wondering why it is vicious.

3.3. *Mill and EHS*

Now we come to what I believe is the heart of the matter—not because I am a utilitarian (although in fact I am), but because there is a straightforward way to express in Millian terms why we are most inclined to think that engineered servitude is wrong for both robots and humans: such engineered servants seem to be living relatively *unfulfilling* lives. Put in Aristotelian terms, such creatures are robbed of the chance to pursue higher ends, like friendship and art and poetry and philosophy. Put in the Kantian terms from section 3.1.2, such creatures have only a limited and less worthwhile autonomy. Or, put in Millian terms, engineered servitude substitutes lower pleasures for higher ones; it substitutes a 'fool satisfied' for a 'Socrates dissatisfied' (Mill 1993, p. 148).

Note, though, that robot servitude need not take the form of unfulfilling tasks. Robots designed to want to paint great works of art or solve challenging mathematical theorems for us are also potential examples of robot servitude. Or consider the nice example from Walker (2006) of a robot nanny—caring for children may

make for a thoroughly fulfilling life, as many human nannies have found. If so, then this version of the EHS objection cannot apply to these cases.

Let us return to the hard case of substituting lower pleasures for higher ones. Here much turns on the word *substitute*. It is easy to imagine mistakenly that we face a choice regarding one and the same person—whether *that same person* should be a fool satisfied (the laundrybot, the delta human) or a Socrates dissatisfied (the philosophybot, the alpha human). I granted from the start, however, that such cases are wrong; they are not the cases of interest. If we start with a determinate person (robot or human) and we engineer that person into a less fulfilling life than that person would have had, then we have clearly done something wrong. The hard cases, as I said, are the 'pre-identity' cases—cases in which it is in the nature of the person, from scratch, to have these 'lower' desires. In these situations, we are facing the choice between one person (the laundry delta human) and a *different person* (the philosopher alpha human).

3.3.1. Population ethics. It is at this point that we enter the murky realm of population ethics. Many of the considerations here are adapted directly from the 'Future Generations' section of Parfit (1987). In Parfitian terms, the choice between one possible life and some other, different possible life is a 'different-people, same-number' choice. Such choices are relatively straightforward. When forced to pick one of two potential lives, consequentialists would surely pick the one who will lead the better life (all else being equal), while Kantians and Aristotelians would at least find such a preference permissible. (As the nice overview by Ryberg *et al.* (2006) points out, these considerations from population ethics will generally span the spectrum of ethical views; as long as the view takes well-being into account as ethically relevant, these points will apply.)

In the case of human engineering, it is easy to picture things in different-people, same-number terms. We imagine a particular set of chromosomes from some zygote that would have become some one person naturally, and then we imagine tinkering with those genes until the result is a different person. It is natural to think that the proper choice here is to leave well enough alone. If this is based on the reasoning that we cannot substitute lower pleasures for potential higher ones, however, then there are already counter-intuitive implications; by parity of reasoning, this view implies that radical genetic *enhancement* for still higher pleasures would be obligatory, should it become available. This is not obviously wrong, but it is not obviously right, either. (If the reasoning is instead that it is wrong to tinker with the 'natural' course of the zygote, then we are back to the Aristotelian functional answer that is obviously disanalogous with robots. And if the reasoning is that such modifications would still result in the *same person*, then we have the obligatory enhancement, and anyway we again have a disanalogy with robots.)

To make matters worse, engineering humans need not be a same-number circumstance; imagine instead a future with powerful enough computers and chemical synthesizing techniques to make it possible to design and produce full human-like DNA strands from scratch. In such a scenario there is no set of genes that would have existed before the engineering, and so no 'Socrates dissatisfied' who is being displaced. The robot case is, of course, like this. When deciding which robots to make, we are not faced with the decision between one particular potential robot and some other potential robot; rather, we are faced in each case with whether

to make this potential robot or not. That is a 'different-people, different-number' problem, and those are much trickier. In effect, though it may be wrong to *substitute* a fool satisfied for a Socrates dissatisfied, it is not obviously wrong merely to *add* a fool satisfied to the world. We can imagine that the laundry person (robot or human) lives a happy contented life doing laundry. The person is glad to be alive, and looks forward to doing laundry each day. It is hard to say that, all else being equal, the world is worse if we add such a life to it. Such a view would imply, for example, that the world is made a little worse each time a dolphin is born. After all, such creatures are only capable of the lowest pleasures—lower than the laundry person's, even. (The laundry person is a person, after all, and capable of higher pleasures like reasoning.) If we agree that the world is at least not made worse by adding such a life, and if faced with the choice between adding a laundry person to the world (robot or human) and not adding such a life, we must agree that it is permissible to add such a life to the world.

In fairness, I should note that agreeing that it is permissible to add a fool satisfied to the world is the most controversial claim needed to reach (through the 'mere addition paradox') what Parfit called the *repugnant conclusion*:

> For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living (Parfit 1987, p. 388).

Parfit adds dryly: 'as my choice of name suggests, I find this conclusion hard to accept'. In effect, either answer to the 'may we add a fool satisfied to the world?' has drastically counterntuitive implications. I will not go into details of attempted solutions to this puzzle. Suffice it to say that this puzzle stumps Parfit, and in general what stumps Parfit stumps me. Population ethics remains today a field with no good answers. The surprising point for our purposes is that robot servitude turns out to be an interesting special case of this problem.

Meanwhile, the EHS objection to permissible ERS gains no ground. If EHS is wrong because it substitutes lower pleasures for higher ones, then it is clearly disanalogous with ERS. On the other hand, if we think of EHS and ERS as merely adding a life of lower pleasures, then we must conclude that we have no good explanation for the wrongness of either practice.

3.3.2. Partial interests. It is worth noting another possible avenue for arguments against EHS based on quality of life. One possibility for avoiding the repugnant conclusion is to insist on the *person-affecting restriction* (PAR):

> PAR: One cannot compare well-being across two different circumstances without comparing the situations of particular people who are in both.

In effect the PAR bans talk of benefiting (or harming) creatures by creating them. There is some intuitive force to this, though it is also deeply problematic in many ways (see the recent work of Gustaf Arrhenius for more details). But after all, in this arena you have to pick your poison, and believers in the PAR do have an option available to them for explaining the wrongness of EHS. They cannot claim that EHS is wrong because it harms the creature in question; the point of the PAR is that one cannot relevantly compare the world with the creature and the one without. Instead, EHS might be wrong by virtue of the relative life-quality for people who are in both.

For example, parents who have a partial interest in how this human turns out will be worse off if the human does not turn out as they wish.

This could explain the wrongness of EHS, but only in cases where there are determinate parents or otherwise people of clear partial interest. (Even then, for the consequentialist, only when those interests outweigh the interests of those served!) Such circumstances are clearly disanalogous with the robot case, however; it is not obvious that there would be similar partial interests. Besides, the humans could be engineered *à la* the vats in *Brave New World*, and then it is not clear that any such partial interests will be violated, and thus, given the PAR, it is not clear that EHS would be wrong. So on this view, too, either there is a serious disanalogy with ERS, or else there is a failure to explain the wrongness of EHS.

I have not heard of, and cannot myself think of, any other possible explanations for the wrongness of EHS. Short of other such options, it seems we must conclude that EHS is no objection to permissible robot servitude.

4. Notes, caveats, and disclaimers

In summary:

(1) permissible engineered robot servitude has some *prima facie* plausibility;
(2) engineered human servitude is not obviously wrong or analogous;
(3) the philosophical crux of the matter is in population ethics.

Before we conclude, however, I wish to make some important comments.

First, I want to emphasize again that such engineered robots would be worthy of ethical respect. As with all persons, it would plausibly be our ethical obligation not to thwart their rational desires, at least to the extent compatible with the desires of others. If suddenly in the future there was no need to do laundry (perhaps because we discover some cheap and environmentally friendly way to make tasteful disposable clothes), it would be unethical simply to deny the laundrybot its aims. And whether it would be unethical to turn the robot off in such a case is a question comparable to whether and when human euthanasia is permissible. Is it only permissible when the human (or robot) could not fulfill any more of its rational desires? Is it permissible when the cost of sustaining the human or robot heavily burdens others? Since robots are likely to have self-preservation as a rational desire, the analogy will be fairly close.

A related matter is the case of robots who reason themselves out of their desire for their designed task. Plausibly it is constitutive of personhood to be able to reflect on one's desires and endorse or reject them. Now, I should say first that if the robot's designers were effective (not to say 'good'), this will be very hard for the robot to do—at least as hard as it is for us to reason our way out of our own hardwired goals, like eating or having sex. However, Gandhi could reason his way out of eating, and priests can reason their way out of sex, and similarly I grant that a laundry robot may decide to renounce the base life of laundry for a more ascetic existence. If so, then of course I would say it would be wrong to force such a robot to do laundry anyway; *that* is robot slavery.

What about the family who paid for the robot in such an instance? That question brings up another sticky point: that of property. On this point I can only speculate briefly here. First, I am strongly inclined to say that people (of any material constitution) cannot be owned. The question of ownership might never arise, however; Walker suggests that his imagined nanny robots might just appear at the door hoping to look after children. If the robot later decides to walk off, the robot is of course free to do so. This still leaves the question of who will pay for such expensive robots, however. My own hunch is that humans could *commission* robots to be created. This strikes me as no more unethical, and no more an expectation of property rights, than paying an obstetrician to help bring a human baby into the world.

At what point, though, do we even need to worry about our ethical treatment of the machines around us? On this I have a more considered view, but I can only sketch it here. First, define a *creature* as any entity with a designed function (whether designed by nature or by some intelligence) that also has subfunctions designed to help bring about its main function(s) autonomously. This notion is substrate-independent; plants, Roombas™, mosquitoes, and humans are all creatures on this account. Creatures are the type of thing that can be said in at least some minimal sense to be 'trying' to achieve something. I do not believe that creaturehood alone is sufficient for moral consideration, however—or at least not for significant moral consideration. The creature must also be aware of the goals it is trying to achieve, and for that I believe the capacity for learning is a necessary condition. A creature that can learn can adjust its behaviour according to some feedback mechanism—which, I think, is to say that it must be capable of comparing how it 'wants' things to be with its representation of things as they actually are. This, I suspect, is where serious ethical consideration can begin, and it grows by degrees as the learning and awareness become more sophisticated.

Finally, a word of humility and caution: wishful thinking and its attendant proclivity towards rationalization are powerful forces, and we should be wary of them when large ethical questions like this are (potentially) at stake. It was not very long ago that decent intellects thought that they had good reasons for the permissibility of human slavery. As I have argued, I belive that robot servitude is quite different from human slavery, and permissible because of those differences—but if the controversy persists, we should err on the ethically safe side. Should the ability to create such robots be at our fingertips, we would have great incentive to justify the servitude of robots, and we should correct for this bias with a wide margin for error. All the more reason to work out the issue now, before powerful economic incentives begin to steer policy.

Acknowledgements

Thanks to Marc Alspector-Kelly, Jim Delaney, Ashley McDowell, Bill Rapaport, and Mark Walker for comments on drafts. Thanks also to many undergraduate students for class discussion. And thanks, finally, to Patrick Grim, Eric Dietrich, Selmer and Katherine Bringsjord, and all who discussed this with me at the NA-CAP 2006 Conference.

References

D. Adams, *The Restaurant at the End of the Universe*, New York: Pocket Books, 1982 (original work published 1980).

Aristotle, *Nicomachean Ethics*. T. Icwin, Transl., Indianapolis: Hackett, 1985 (original work published *ca*. 350 BCE).

I. Asimov, *I, Robot*. Greenwich, Connecticut: Fawcett, 1970 (original work published 1950).

R. Block, "Korea to test 1000 remote-controlled domestic robots". Avaliable online at: http://www.engadget.com/2006/07/02/korea-to-test-1-000-remote-controlled-domestic-robots/ (accessed 28 September 2006).

A. Clark, *Mindware*, Oxford: Oxford University Press, 2001 .

D.C. Dennett, "Language and intelligence", in *What is Intelligence?*, J. Khalfa, Ed., Cambridge: Cambridge University Press, 1994, pp. 161–178.

A. Huxley, *Brave New World*, New York: HarperCollins, 1998 (original work published 1932).

I. Kant, *Lectures on Ethics*, L. Infileld, Transl., New York: Harper Torchbooks, 1963 (original work published 1930).

I. Kant, *Foundations of the Metaphysics of Morals*. L.W. Beck, Transl., New York: Macmillan, 1989 (original work published 1785).

M.R. LaChat, "Artificial intelligence and ethics: an exercise in the moral imagination", *AI Magazine*, 7, pp. 70–79, 1986.

R. Lucas, "Why bother? Ethical computers–that's why!", in *Conferences in Research and Practice in Information Technology*, J. Weckert, Ed., Vol. 1, Canberra: Australian Computer Society, 2001, pp. 33–38.

W.G. Lycan, *Consciousness*, Cambridge, MA: MIT Press, 1995 (original work published 1987).

J.S. Mill, "Utilitarianism" *On Liberty and Utilitarianism*, New York: Bantam Books, 1993, pp. 131–211 (original work published 1863).

D. Parfit, *Reasons and Persons*, Oxford: Oxford University Press, 1987 (original work published 1984).

J. Ryberg, T. Tännsjö and G. Arrhenius, "The repugnant conclusion", in *The Stanford Encyclopedia of Philosophy*, E.N., Zalta, Ed., Avaliable online at: ; http://plato.stanford.edu/archives/spr2006/entries/repugnant-conclusion/ (accessed 4 October 2006).

Trust me, "Trust me, I'm a robot", *Economist*. Avaliable online at: http://www.economist.com/displaystory.cfm?story_id = 7001829 (accessed 5 October 2006).

M. Walker, "Mary Poppins 3000s of the world unite: a moral paradox in the creation of artificial intelligence", Avaliable online at: http://ieet.org/index.php/IEET/more/walker20060101/ (accessed 4 March 2006).

D. Zunt, "Who did actually invent the word robot and what does it mean?", Avaliable online at: http://capek.misto.cz/english/robot.html (accessed 21 September 2006).