18 Designing People to Serve

Steve Petersen

Fiction involving robots almost universally plays on a deep tension between the fantasy of having intelligent robot servants to do our every bidding, and the guilt over the more or less explicit possibility that having such intelligent creatures do our dirty work would simply be a new form of slavery. The robot character is frequently a sympathetic antihero who gets mistreated by its callous, carbon-chauvinist human oppressors. Our guilt over such a scenario often manifests as a fear that the robots' servile misery would drive them to a violent and relentlessly successful uprising. As commonly noted, the very word "robot" has its roots in just this scenario; it first appears in Karel Čapek's play *R.U.R: Rossum's Universal Robots*, in which a brave new world of robot servants eventually rebel against their oppressive human masters. Čapek chose the word "robot" to invoke the Czech word *robota*, which translates to "drudgery" or "forced labor."[1] Čapek's play seems to have set the stage for a very long list of books, movies, and television shows about robots to follow. Try to list a few robot stories that *don't* fit this fantasy-guilt complex, and I'm confident you will generate a sizable list of examples that do fit it yourself.

So this aspect of robot ethics has long been in our culture—but it is only just beginning to appear in academia. The few authors who directly confront the ethics of robot servitude tend to conclude one of three things. Some propose that such robots could never be ethical subjects, and so we could not wrong them in making them work for us any more than we now wrong a washing machine. Others agree that robots could not be of ethical significance, but say we must treat them as if they were anyway, for our own sake. Still others conclude that robots *could* someday have genuine ethical significance similar to ours, and that therefore it would be unethical for them to perform menial tasks for us; it would simply be a new form of slavery.[2]

My own position, originally developed in Petersen 2007, is quite different from all of these. First of all, I do think it is possible to create robots of ethical significance—even to create *artificial people*, or *APs* for short. In a tradition with its roots in John Locke, philosophers tend to distinguish the biological category *human* from the much more philosophically rich category *person* ([Locke 1690] 1838, II.xxvii). To say that

something artificial could be a person is to say in part at least that it could have full ethical standing like our own. On this usage, for example, ET the Extra-Terrestrial would be a person, but not a human. ET does not share our DNA, but this is irrelevant to his ethical standing; he is as ethically valuable as we are. In other words, to be a person does not seem to require being made of the particular material that happens to constitute humans; instead, philosophers tend to agree, it requires complicated organizational patterns that the material happens to realize. And thus, assuming we could eventually make a robot who has the same relevant complicated organizational patterns that we and ET have, then that robot would also be a person—an artificial one.

I *also* think that although such robots would be full-blown people, it might still be ethical to commission them for performing tasks that we find tiresome or downright unpleasant. There can, in other words, be artifacts that (1) are people in every relevant sense, (2) comply with our intentions for them to be our dedicated servants, and (3) are not thereby being wronged. I grant this combination is prima facie implausible, but there are surprisingly good arguments in its favor. In a nutshell, I think the combination is possible because APs could have hardwired desires radically different from our own. Thanks to the design of evolution, we humans get our reward rush of neurotransmitters from consuming a fine meal, or consummating a fine romance—or, less cynically perhaps, from cajoling an infant into a smile. If we are clever we could design APs to get their comparable reward rush instead from the look and smell of freshly cleaned and folded laundry, or from driving passengers on safe and efficient routes to specified destinations, or from overseeing a well-maintained and environmentally friendly sewage facility. After all, there is nothing intrinsically unpleasant about hydrogen sulfide molecules, any more than there is anything intrinsically pleasant about glucose molecules. The former's smell is aversive and the latter's taste is appetitive *for humans*; APs could feel quite differently.[3] It is hard to find anything wrong with bringing about such APs and letting them freely pursue their passions, even if those pursuits happen to serve us. This is the kind of robot servitude I have in mind, at any rate; if your conception of *servitude* requires some component of unpleasantness for the servant, then I can only say that is not the sense I wish to defend.

18.1 The Person-o-Matic

To dramatize the ethical questions that APs entail, imagine we sit before a *Person-o-Matic* machine. This machine can make an artificial person to just about any specifications with the push of a button. The machine can build a person out of metal and plastic—a robotic person—with a circuit designer and an attached factory. Or, if we wish, the machine can also build a person out of biomolecules, by synthesizing carefully sequenced human-like DNA from amino acids, placing it in a homegrown

cellular container, and allowing the result to gestate in an artificial uterus. It can make either such type of person with any of a wide range of possible hardwired appetites and aversions.[4] Which buttons on the Person-o-Matic would it be permissible to press?

It may be difficult to reconcile ourselves to the notion that we could get a genuine *person* just by pushing a button. My students like to say that nothing so "programmed" could be a person, for example. But—as the carbon-based AP case makes especially vivid—the resulting beings would have to be no more "programmed" than we are.

A more sophisticated version of this complaint is in Steve Torrance's "Organic View" (2007). He argues that only "organic" creatures could have the relevant ethical properties for personhood, and so "artificial person" is practically a contradiction in terms. Of course, a great deal hinges here on just what "organic" means. Torrance seems to use it in three different ways throughout his paper: (1) *carbon-based*, (2) *autopoietic*, and (3) *originally purposeful*. This quotation, for example, illustrates all three: "Purely electrically powered and computationally guided mechanisms [sense 1] . . . cannot be seen, except in rather superficial senses, as having an inherent motivation [sense 3] to realize the conditions of their self-maintenance [sense 2]: rather it is their external makers that realize the conditions of their existence [sense 3]" (Torrance 2007, 512–514). But none of these three senses of *organic* is enough to show that APs are impossible.

Consider first the sense in which it means *carbon-based*. Torrance provides no argument that only carbon could ground ethical properties; indeed, philosophical consensus is otherwise, as mentioned earlier. Besides, even if people do have to be organic in this sense, APs are still possible—as Torrance acknowledges (2007, 496, 503)—because it is in principle possible to create people by custom building DNA.

Torrance officially uses *organic* in the second sense, to mean *autopoietic*. Roughly, something is autopoietic if it can self-organize and self-maintain. But this is a purely functional notion; there is no reason inorganic compounds couldn't form something autopoietic. Indeed, the well-established movement of situated, embodied, and embedded robotics emphasizes getting intelligence out of just such lifelike properties.[5] Rodney Brooks's Roomba, for example, avoids treacherous stairs and seeks its power source after a long day of vacuuming. Such robots already have rudimentary self-organization and self-maintenance.

Lurking behind the criterion of autopoiesis is the third sense of *organic*, and what I suspect to be the core of the matter for Torrance's argument: the presence of *inherent function* or purpose. Torrance is claiming, in effect, that when something gains a function by another's design, the function is not inherent to that thing, and so it is not "original." And, Torrance seems to hold, only original functionality can ground ethical value. In other words, just in virtue of resulting from another's design, a thing cannot be a person. (Perhaps this is what my students mean by something being "programmed.")

E1

If correct, this would by definition rule out all APs, carbon-based or not. But, aside from having scant motivation, it proves too much. By this criterion, if traditional Christian creationism proved true and God designed us, then we humans would not be "organic" either, and so not people. I'm strongly inclined to agree that evolution, and not God, designed humans—but it would be very odd if our ethical standing were so hostage to the truth of this claim. For another example closer to home, it seems that our biological parents count as our "external makers," who were moved to "realize the conditions of [our] existence" (though probably not in a traditional laboratory setting). Despite such external makers, we manage to have the properties required to be people.

Finally, consider Labrador Retrievers. They are not people, of course, but they do have ethical standing, and they were deliberately designed, via artificial selection, to enjoy fetching things. Does this mean that they have no "inherent motivation" to fetch? Anyone who has spent time with a retriever can see that the dog, itself, wants to fetch—whatever the source of that desire. Furthermore, satisfying this desire is part of the well-being *for that dog*, even though that desire was designed by intelligent outsiders. Similarly, we did not give ourselves all our desires; some of them, such as for food, are just plain hardwired. It is hard to see why ends given by evolution are "original," but ends given by the design of an intelligence are not. In both cases, there is a very natural sense where our ends seem plainly derivative.

Still, I think Torrance is onto something important; in fact, I agree that for something to be intelligent, autopoietic, and a subject of ethical value, it must have a function *for itself*. Teleology is a notorious can of worms in philosophy, and can hardly be settled here. For our purposes, we just need the claim that one way for something to get a function for itself—an "original teleology"—is from the design of another intelligence.

So now perhaps we are in a position to agree that pushing a Person-o-Matic button would result in a real person of intelligence and ethical value, comparable to our own. When we picture this vividly, I think typical intuitions incline us to say that pushing few, if any, of the buttons is permissible. The case is so far removed from our experience, though, that it is hard to trust these intuitions—especially since there are good arguments that say it *is* permissible to press quite a few of them.

18.2 The "Typical" Person Case

Suppose first you notice buttons for building an organic person, just like you (presumably) are. (From here I will use *organic* just to mean *carbon-based*.) Perhaps, after you feed it the complete information about your DNA makeup and the DNA makeup of a willing partner, the Person-o-Matic uses a combination of this information to construct from scratch a viable zygote that matures in an artificial uterus, much later

producing an infant, exactly as might have been produced by the more traditional route. Here we leave a great deal of the design up to chance, of course; our intention is not to create a servant, but roughly just for the Person-o-Matic to build a new human, or anyway a human-like person.[6] The scenario may be intuitively distasteful or even abhorrent, but it is very hard to give reasons for why creating such a person would be *wrong*. After all, it results in people just like the people we now create by traditional means. There may be circumstances in which just the creating of a new person is unethical, of course—due to overpopulation or some such—but that would hardly be unique to APs. If anything is uniquely wrong about this case, then, it must be in the *method* for creating the person, rather than the outcome. But even the method seems no less ethical than a combination of in vitro fertilization, artificial implantation, surrogate mothers, and a host of other techniques for creating people that are already in practice. No doubt bioethics is another can of philosophical worms, but the case at hand here is not so different from bioethical cans already wide open. Indeed, using the Person-o-Matic this way could plausibly bring ethical benefit to a great many couples who are not otherwise able to have biological children.

Probably, the most natural way to express our intuitions against the permissibility of this case is to say that such a procedure for making a person like us would be "unnatural." This word shows up frequently when people are confronted with new technology. As a clever novelist once put it:

> 1. Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works.
> 2. Anything that's invented between when you're fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it.
> 3. Anything invented after you're thirty-five is against the natural order of things. (Adams 2002, 95)

The point, of course, is that much of what we consider "natural" today may have looked horrifyingly unnatural to those just a generation or two behind us. To say "unnatural" in this way just means "new enough to make us wary and uncomfortable." When the word means only this, it has no philosophical weight. Our gut reactions are often wise for being wary of the new and strange, but rejecting something *because* it is new and strange is quite different. We do not now consider flying, cell phones, radiation treatment, or artificial hearts wrong because they would have been distressing to those before us.

It seems then that it is hard to explain why it would be wrong to push such a button. As it happens, though, next to those buttons is another row of buttons that offer the option to create a person much like us, except inorganic—a robot. Aside from desires and goals that are particular to the material makeup, we can suppose the robot is designed to have hardwired interests very like ours, and will also be strongly influenced in a unique way by its educational environment just as we were. Would it be

wrong to push any of those buttons? It seems there are only a few avenues for trying to explain such wrongness. One is to say that though the resulting person would be like us in all relevant mental respects, just the fact of its different material constitution makes its creation wrong. Another might be that the desires unique to our organic constitution are relevant—that, for example, it is okay to make an AP who likes to consume carbohydrates, but not one who likes to consume pure hydrogen. I trust neither of these avenues looks very promising. If not, and short of other explanations of asymmetry, the organic and the inorganic cases seem to be morally equivalent. We must conclude that making a robot with predispositions like ours is no more wrong than having a biological child would be.

18.3 The "Enhanced" Person Case

We next notice a bank of buttons to create organic people who are still very much like us, but who have been "enhanced" in any of various ways. Some buttons offer to design the person so that she is immune to common diseases. Of more interest for us, some buttons offer to alter the person's hardwired desires—so that perhaps she is also immune to the lures of tobacco, or enjoys eating healthy greens more than usual. Other buttons offer to tailor more abstract desires, so that, for example, the AP gains greater intrinsic pleasure than typical from pursuits we consider noble, such as sculpting or mathematics. Would it be wrong to press a button to bring about this type of person?

Again, despite what qualms we might have, it is hard to say why it would be. Given that parents and mentors expend great and generally laudable effort on the nurture side to bring about such results, it is at least a bit odd to say that bringing them about from the nature side would be wrong.

Probably the best argument against creating the "enhanced" person suggests we have robbed the resulting person of important autonomy by engineering such desires. On this view, it is one thing to encourage such desires during the person's upbringing, and another to hardwire them ahead of time. Of course, free will is yet another philosophical can of worms, and one into which we can only peek here—but again, it is a can of worms that is already open, and hardly unique to APs. Some humans are now naturally born with stronger resistance to tobacco's appeal, for example, and it may well be that some are naturally born with stronger predilections for math or art. At any rate, we all come into existence with hardwired desires, and whether they are "enhanced" or "typical" does not seem relevant to whether they are enacted freely.

Imagine, for example, that way down the road—perhaps hundreds of millions of years later—natural selection has shaped humans so that they no longer enjoy tobacco, and they are born with a random mix of significantly stronger desires to do art or

science or other lofty pursuits. This seems possible at any rate, and it would be very odd to say that those future humans would thereby have less autonomy than we have. But our Person-o-Matic can now make a molecular duplicate of such a future possible person. If the future product of natural selection is free and the duplicate AP is not, then one's autonomy depends on how one is brought into existence, even if the result is otherwise exactly the same. It is to say, in effect, that intelligent design does not create an *original* function after all.

I have already argued against this position; I hope, on reflection, it is hard to endorse. It is more interesting to examine what tempts us into this view in the first place. Perhaps, it is simply the familiar queasiness of the "unnatural." Another possibility is that we confuse the case at hand with a more familiar one: that of brainwashing a person with contrary desires already in place.

Another possible source of confusion is in the imagined relative *strength* of these inclinations. Perhaps typical people are free, despite being born with strong dispositions because, we think, they are still able in principle to resist them. Whatever this "ability" amounts to, though, we can suppose APs have it, too. It is plausibly a necessary condition of personhood that one be able to reflect on one's desires, for example, and reconsider them (Frankfurt 1971). An enhanced AP might crave mathematics or sculpting as much as a typical human craves food. But Gandhi could reason himself out of acting on his food craving, and the enhanced AP might similarly reason herself out of her cravings, because she is a person able to reflect on them. So, the AP seems to be as free as we humans are—however free that might be—and the objection from autonomy fails.

It is no great surprise when we see another row of buttons on the Person-o-Matic for creating enhanced APs that are inorganic. These buttons result in robots who love to carve elegant statues or prove elegant theorems. Again, pushing these buttons seems morally equivalent to the ones for the organic APs. If so, then creating a robot who loves to pursue art or science is no more wrong than giving birth to a human who gained the same predispositions through natural selection.

Notice, though, that pushing buttons in either of these rows is already at least tantamount to designed servitude. Suppose we commission an AP who is very strongly inclined to help find a cure for cancer. Is this AP our willing servant? If so, then I have already shown that we can design people to serve us without thereby wronging them.

18.4 The "General Servitude" Case

A scientist dedicated to curing cancer, even as a result of others' desires, may not seem like a clear case of servitude. Clearer cases follow readily, though—because one enhancement for a person, plausibly, is general beneficence. Sure enough, a prominent button on the Person-o-Matic designs an organic person who gains great pleasure E1

simply from bringing about happiness in other people. The AP who results genuinely likes nothing more than to do good and will seek opportunities to help others as eagerly as we seek our own favorite pleasures.

Again, it seems possible that natural selection could bring about humans like this in the far future—if group selection turns out to be a force for genetic change after all, for example—and it would not then be wrong to give birth to one. (Indeed, it sounds like a pretty good world into which to be born.) Again, the Person-o-Matic could create a molecular duplicate of such a person. Again, it is hard to see why the naturally selected person would be permissible and not the intelligently designed one. Again, it does not matter, on ethical grounds, whether the resulting AP is organic or inorganic. So, again, we have to conclude that commissioning a robot who wants to help people above all else is no more wrong than giving birth to a human who gained such beneficence through natural selection. The resulting APs would behave much as though they were following Isaac Asimov's Three Laws of Robotics from his *I, Robot* series ([1950] 1970)—except they would also be helpful to other APs. And this time it seems very clear that the resulting AP would be a dedicated servant to the people around it.

18.5 The "Specific Servitude" Case

Closer still to the *I, Robot* scenario are APs who are designed not to seek the happiness of people generally, but rather the happiness of humans specifically. This is a more task-specific kind of servitude. Still, more specifically, perhaps they are designed to seek the health and well-being of human children—or even your particular children, as Walker pictures his *Mary Poppins 3000*:

> 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 utmost haste and begs you for the opportunity to look after your children. (2006)

In fact, the first robot we meet in the *I, Robot* stories is a similar nanny. Inspection of the Person-o-Matic of course reveals "nanny" buttons, as well as buttons that engineer people to derive great joy out of freshly cleaned and folded laundry, or from driving safe and efficient routes to specified destinations, or from clean and efficient sewers. These buttons are probably the most controversial ones to push; they evoke the gruesome "delta caste" of people engineered for menial labor in Aldous Huxley's *Brave New World* ([1932] 1998)—especially in the case of organic, human-like APs.[7] Though surely our intuitions rebel against these cases most of all, it is surprisingly

difficult to find principled reasons against pushing even these buttons. The three best of which I know are:

1. The resulting AP would have impermissibly limited autonomy.
2. The resulting AP would lead a relatively unfulfilling life.
3. The resulting AP would desensitize us to genuine sacrifices from others.

I will address each reason separately.

18.5.1 Specific Servitude and Autonomy

First, consider the objection from autonomy. Walker, for example, says that in making one of his imagined robot nannies we have just made a "happy slave," because "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" (2006).

I have already addressed the autonomy argument in some detail for the enhanced person case. Those arguments carry over to this case at least to the extent that the content of one's hardwired desires are irrelevant to the autonomy with which they are pursued. If one AP is made with a strong desire to sculpt, another with an equally strong desire to look after your children, and yet another with an equally strong desire to do laundry, then it seems they should all be equally free. If we object to making one and not the other, then it does not seem to be on *autonomy* grounds.

We are more tempted here than in the "enhanced" person case to object from autonomy, though, and I can think of two reasons why: first, it is harder for us to conceive of a person who genuinely wishes such ends for themselves, at least without our coercing them from other, more "natural" desires. Second, the desired ends these APs seek serve us in a much more obvious way. This combination has the effect of convincing us that the APs are being used as a mere means to our ends—and according to a flourishing ethical tradition founded by Immanuel Kant, it is an impermissible violation of autonomy to use any person as a mere means to an end ([1785] 1989).

The "mere" use as means here is crucial. In your reading this chapter, I can use you as a means to my ends—which may be your finding the truth of some difficult ethical claims, or sharing my philosophical thoughts, or my gaining philosophical glory and tenure. Meanwhile you can use me as means to your ends—which may be your gaining a wider perspective on robot ethics, or entertaining yourself with outlandish views, or proving me wrong for your own philosophical glory. This is permissible because we are simultaneously respecting each other's ends. And here, of course, we see that the same is true of the task-specific APs: though they are a means to our ends of clean laundry and the like, they are simultaneously pursuing their own permissible ends in the process. They therefore are not being used as a *mere* means, and this makes all the ethical difference. By hypothesis, they want to do these things, and we are happy to let them.

E1

Now as genuine people, we are supposing these APs are worthy of full ethical respect, and for the Kantian this means supposing they have a required autonomy. This plausibly means, as noted earlier, that such APs are capable of reasoning themselves out of their predisposed inclinations. But first, this could be roughly as unlikely as our reasoning ourselves out of eating and sex, given the great pleasure the APs derive from their tasks. Second, if they should so reason, then of course I would not defend making them do their tasks anyway; that would be wrong on just about any plausible ethical view.[8] Indeed, if the APs do not reason themselves out of their joy in washing laundry, to give an example, and if suddenly there were no more laundry to do—perhaps because nudity became the fashion—it would be our obligation to help them out by providing them with some unnecessarily dirty clothes.

18.5.2 Specific Servitude and a Fulfilling Life

Perhaps what's behind the autonomy objection is that, despite the fact that the AP comes into existence with these desires, that AP was still "coerced" into an otherwise aversive task. In other words, it is really about the content of the desires—just to bring the APs into existence with such abject desires is to manipulate them unfairly. If so, this is really a form of the next objection: that to create a being who enjoys pursuing such menial tasks is to create someone who we know will live a relatively unfulfilling life, and this is impermissible.

First of all, it is not obvious that such a life is truly "unfulfilling." Assuming that the laundry AP deeply desires to do laundry, and has an ample supply of laundry to do, the life seems to be a pretty good one. We should be careful not to assume the AP must somewhere deep down be discontent with such work, just because we humans might be. And though perhaps clean laundry does not seem so meaningful an achievement in the big picture of things, in the *big* picture I am sorry to say that none of our own aspirations seem to fare any better.

Probably the best way to push the objection from an unfulfilling life is through a distinction that goes back to the utilitarian John Stuart Mill: that between "higher" and "lower" pleasures. Mill says that "there is no known Epicurean theory of life which does not assign to the pleasures of the intellect, of the feelings and imagination, and of the moral sentiments, a much higher value as pleasures than to those of mere sensation" (1871, 14).

As he famously summarizes, "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied" (Mill 1871, 14). Perhaps the task-specific AP is merely a "fool satisfied."

If a strong, hardwired reinforcement for some achievement is sufficient for it to be a lower pleasure of "mere sensation," then even an AP designed with Socrates' taste for philosophy is only living the life of a fool satisfied. Such a criterion for higher and lower pleasures seems arbitrary. If instead we take the higher pleasures to be, as Mill

insists, simply what the person who has experienced both will prefer, then it seems they will be highly dependent on the person and their own tastes.[9] If so, then the AP, with quite different interests from ours, might well prefer laundry over a good production of Shakespeare, even after experiencing both—and so laundry may count as that AP's higher pleasure. If experiencing higher pleasures is, in turn, what constitutes a fulfilling life, then that AP is leading a fulfilling life by doing laundry.

Suppose we grant, though, that for any person of whatever design, doing laundry is not as fulfilling as (for example) contemplation or artistic expression. Even under this assumption, it is still not obvious that it would be wrong to commission such APs.

For one thing, there is no principled reason the AP could not pursue both types of pleasure; we humans manage it, after all. We tend to seek out and enjoy the higher pleasures only after an adequate number of the lower ones have been satisfied, and this fact does not make our lives unfulfilling. And even if given the opportunity to indulge in the lower pleasures exclusively, many of us (who have experienced both) will get bored and seek the higher ones, at least for a while. The APs could well be similar, especially if we design them so; perhaps after bingeing on their baser desires for washing laundry, the sated APs will then turn to Shakespeare or Mahler for a while.

Suppose, though, that the AP spends its whole life cheerfully doing laundry—perhaps at a large twenty-four-hour facility, rather than in a family's home—without ever experiencing what we are supposing to be higher pleasures. Here, surely we have a case of the "fool satisfied." And, the claim goes, bringing about such a life is wrong, because it is not as good as the life of a Socrates dissatisfied.

Here is a dizzying question, though: who exactly is wronged by pushing the button for a laundry AP? It cannot be the resulting laundry AP, because any time before the AP's desires existed is also a time before the AP existed, and so there was no person being harmed by their endowment. Had we pushed the button for the sculptor AP instead, we would have thereby brought about a *different* person, and so the laundry AP cannot benefit from our pushing the sculptor AP button.[10]

A similar case can be made that the miller's daughter was not wrong to promise her firstborn to Rumpelstiltskin, since had she not done so she never would have married the king, and a different first child would have been born to her—if any. Therefore, assuming the child sold into Rumpelstiltskin's care would rather have that life than no life at all, the promise could hurt no one, and so is not wrong. This is surely counterintuitive.

Ethicists will recognize this as what has come to be called *the nonidentity problem*.[11] This problem is a part of *population ethics*—yet another philosophical can of worms worth more attention than I can give it here. (This abundance of nearby philosophical worms is, for me, part of the topic's appeal.) According to a plausible answer to the puzzle already discussed, though, it is better from an ethical standpoint to bring about

the sculptor AP than the laundry AP, despite the fact that bringing about the laundry AP instead would harm no one in particular. In other words, an act can be wrong even if it harms no one person, just because it causes less overall well-being than alternatives.[12]

Thus, we might agree that choosing the laundry AP button over the sculptor AP button is wrong, when given the opportunity. But suppose the choice is not exclusive, and you have the opportunity to push *both*. Assuming it is permissible to push the button for the sculptor AP, would it be wrong to push the button for the laundry AP in addition? In this case, we are not substituting a comparatively worse life for a better one; rather, we are simply adding a worthwhile life to the world, even though there are or could be better ones. If this is wrong, then a great deal of our current policies should change drastically. We should prevent the birth of nonhuman animals as best as we are able, for example, since they are capable of only the very lowest pleasures, and so, according to this view, it is wrong to add them to the world. We should also make sure that only those people who can be expected to provide the very best lives—whatever those might be—may have children. And if the Person-o-Matic can make people capable of higher pleasures than that of an ordinary human, then humans should stop reproducing altogether.

If we agree that adding worthwhile but nonideal lives to the world is permissible, however, then it is permissible to push the laundry AP button—even under the questionable assumption that the lives of laundry APs are relatively unfulfilling.

18.5.3 Specific Servitude and Desensitization

One last objection to robotic servitude is what I like to call the "desensitization" objection: that having APs do work for us will condition us to be callous toward other people, artificial or not, who do *not* wish to do our dirty work. As David Levy puts it, "Treating robots in ethically suspect ways will send the message that it is acceptable to treat humans in the same ethically suspect ways" (2009, 215).

Those who hold this view generally do not believe that the robots in question are people; they hold that the robots lack some necessary property for ethical value, such as (in Levy's case) sentience.[13] In this form, the objection does not apply to our cases of interest. We should treat APs well, whether organic or inorganic, not because they could be mistaken for people, but because they *are* people. And treating them well—respecting their ends, encouraging their flourishing—could involve permitting them to do laundry. It is not ordinarily cruel or "ethically suspect" to let people do what they want.

Perhaps we can amend the usual desensitization argument to apply to APs, though; perhaps having an AP do laundry for us will condition us blithely to expect such servitude of those who are not so inclined. This argument thus assumes the general population is unable to make coarse-grained distinctions in what different people value. This may well be; humanity has surely displayed stupidity on a par with this

in the past. But we do not normally think that all people like haggis, for example, just because some do, so we seem generally capable of recognizing differences in inclinations. More importantly, the fact that people may make such mistakes is no objection to the position, in principle at least. As Mill said, any ethical standard will "work ill, if we suppose universal idiocy to be conjoined with it" (1871, 35). In this form of the objection, we can respond simply by promising to introduce such APs with caution, and accompanied by a strong education program. As a result, instead of learning that people can be used as means, children might learn about the wide range of ends a person could undertake, and thus gain respect for a more robust value pluralism than they could with ordinary humans alone.

Sometimes this objection rings of a protestant guilt about shirking hard labor. If the concern is that idle hands are the devil's play thing, and that we will grow soft and spoiled with the luxury, then we should also consider whether it is already too late, given the technology we now possess. Not only should we be doing our own laundry, if hard labor is good for its own sake, but we should be doing it in a stream by beating it with rocks.

18.6 Underview

I am not arguing that pushing *any* button on the Person-o-Matic is permissible. For one thing, designing a person who strongly desires to kill or inflict pain would be wrong on just about any ethical view. So would designing a person to lead a predictably miserable life,[14] or to crave tasks that are dangerous for them to do. (With good engineering, though, we can probably make a robot that can *safely* do tasks that are dangerous for humans.)

I am not even sure that pushing the buttons defended above is permissible. Sometimes I can't myself shake the feeling that there is something ethically fishy here. I just do not know if this is irrational intuition—the way we might irrationally fear a transparent bridge we "know" is safe—or the seeds of a better objection. Without that better objection, though, I can't put much weight on the mere feeling. The track record of such gut reactions throughout human history is just too poor, and they seem to work worst when confronted with things not like "us"—due to skin color or religion or sexual orientation or what have you. Strangely enough, the feeling that it would be wrong to push one of the buttons above may be just another instance of the exact same phenomenon.

Notes

1. Zunt (2002) presents a letter of Čapek's in which he credits his brother Josef for the term.

2. For the first view, see Torrance 2007 or Joanna Bryson's less nuanced but provocatively titled "Robots Should Be Slaves" (2010). For the second view, see, for example, Levy 2009; Ronald Arkin

E1

and Mark Walker have also pressed versions of this objection in correspondence with the author. For the last view, see the Walker 2006 and a host of informal online discussions, such as at the American Society for the Prevention of Cruelty to Robots—ASPCR 1999.

3. Compare the intelligent shipboard computer in Douglas Adams's novels, absolutely stumped by why the human would want "the taste of dried leaves boiled in water," with milk "squirted from a cow" ([1980] 1982, 12).

4. The material will of course constrain some of these appetites and aversions. Though philosophers tend to agree that the mental state of *desire* (for example) is a substrate-independent functional role, some particular desires are more substrate independent than others—just as the functional role of a pendulum clock can be realized in wood or brass, but probably not in gaseous helium. See Lycan 1995 for more discussion.

5. They thus practice what Peter Godfrey-Smith calls "methodological continuity" between artificial life and artificial mind (Godfrey-Smith 1996, 320).

6. Perhaps to be part of the biological category *human* requires a certain evolutionary history, so that APs do not count.

7. One extreme thought experiment along these lines is again from the fertile imagination of Douglas Adams: a bovine-type animal designed to want to be eaten, and smart enough to explain this fact to potential customers.

> "I just don't want to eat an animal that's standing there inviting me to," said Arthur. "It's heartless."
>
> "Better than eating an animal that doesn't want to be eaten," said Zaphod.
>
> "That's not the point," Arthur protested. Then he thought about it for a moment. "All right," he said, "maybe it is the point. I don't care, I'm not going to think about it now." ([1980] 1982, 120)

This particular case is probably impermissible on various grounds, however.

8. Since it's become a leitmotif, another example from Adams: "Not unnaturally, many elevators imbued with intelligence . . . became terribly frustrated with the mindless business of going up and down, up and down, experimented briefly with the notion of going sideways, as a sort of existential protest, demanded participation in the decision-making process and finally took to squatting in basements sulking" (Adams [1980] 1982, 47).

9. Mill's test actually insists on the majority of what people would say (1871, 12, 15), but this is even worse; then what counts as a higher pleasure changes depending on how many APs of what type emerge from the Person-o-Matic.

10. One possibility that is probably unique to the inorganic case is when one robot body—humanoid in shape, say—can be programmed either of two ways. In this case, it makes sense to say that particular hunk of material could have been a sculptor or a launderer. If that hunk of material is the AP itself, rather than merely its body, then we can harm *that* AP by pushing the laundry button. But on this account, the AP exists prior to its programming, in that hunk of material. This means it would also harm the AP to, for example, disassemble that body before it ever gets programmed. I take this as a *reductio* of the view that an inorganic AP is identical to its

body, and I leave it to the reader to consider analogies in the organic case. The philosophical problem of *personal identity*—that of determining what changes a person can undergo and still be that same person—is another can of worms beyond this chapter. Suffice it to say that this is not an obviously amenable escape route from the claim on the table: namely, that because no one is harmed by bringing about the laundry AP, it is permissible to do.

11. It is discussed most famously in Parfit [1984] 1987; see Roberts 2009 for an overview.

12. This follows from what Parfit calls the "Impersonal Total Principle."

13. Still, they say, we should treat them well basically for the same reason Kant says we should treat dogs well, even though (in Kant's view) dogs are not subjects of ethical value, either: because "he who is cruel to animals becomes hard also in his dealings with men" ([1930] 1963, 240).

14. More leitmotif: Adams's character Marvin, the "Paranoid Android," was designed by the Sirius Cybernetics Corporation to have the "genuine people personality" of severe depression (Adams [1979] 1981, 93).

References

Adams, D. [1979] 1981. *The Hitchhiker's Guide to the Galaxy*. New York: Pocket Books.

Adams, D. [1980] 1982. *The Restaurant at the End of the Universe*. New York: Pocket Books.

Adams, D. 2002. *The Salmon of Doubt: Hitchhiking the Galaxy One Last Time*. New York: Random House.

Asimov, I. [1950] 1970. *I, Robot*. New York: Fawcett Publications.

ASPCR. 1999. The American Society for the Prevention of Cruelty to Robots website. <http://www.aspcr.com> (accessed April 24, 2010).

Bryson, J. J. 2010. Robots should be slaves. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Y. Wilks, 63–74. Amsterdam: John Benjamins.

Frankfurt, H. G. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68 (1): 5–20.

Godfrey-Smith, P. 1996. Spencer and Dewey on life and mind. In *The Philosophy of Artificial Life*, ed. M. A. Boden, 314–331. Oxford: Oxford University Press.

Huxley, A. [1932] 1998. *Brave New World*. New York: HarperCollins.

Kant, I. [1785] 1989. *Foundations of the Metaphysics of Morals*, trans. L. W. Beck. London: The Library of Liberal Arts.

Kant, I. [1930] 1963. *Lectures on Ethics*, trans. L. Infield. London: Harper Torchbooks.

Levy, D. 2009. The ethical treatment of artificially conscious robots. *International Journal of Social Robotics* 1 (3): 209–216.

Locke, J. [1690] 1838. *An Essay Concerning Human Understanding*. London: Tegg and Co.

Lycan, W. G. 1995. The continuity of levels of nature. In *Consciousness,* 37–48. Cambridge, MA: MIT Press.

Mill, J. S. 1871. *Utilitarianism*, 4th ed. London: Longmans, Green, Reader, and Dyer.

Parfit, D. [1984] 1987. *Reasons and Persons*. Oxford, UK: Oxford University Press.

Petersen, S. 2007. The ethics of robot servitude. *Journal of Experimental and Theoretical Artificial Intelligence* 19 (1): 43–54.

Roberts, M. 2009. The nonidentity problem. In *The Stanford Encyclopedia of Philosophy* (Fall ed.), ed. E. N. Zalta. Metaphysics Research Lab, CSLI, Stanford University. <http://plato.stanford.edu/entries/nonidentity-problem/> (accessed July 14, 2011).

Torrance, S. 2007. Ethics and consciousness in artificial agents. *AI and Society* 22 (4): 495–521.

Walker, M. 2006. *Mary Poppins 3000s of the World Unite: A Moral Paradox in the Creation of Artificial Intelligence*. Institute for Ethics & Emerging Technologies. <http://ieet.org/index.php/IEET/more/walker20060101/> (accessed March 4, 2006).

Zunt, D. 2002. Who did actually invent the word "robot" and what does it mean? <http://capek.misto.cz/english/robot.html> (accessed November 20, 2010).