Robots should be slaves

Joanna J. Bryson

Robots should not be described as persons, nor given legal nor moral responsibility for their actions. Robots are fully owned by us. We determine their goals and behavior, either directly or indirectly through specifying their intelligence or how their intelligence is acquired. In humanising them, we not only further dehumanise real people, but also encourage poor human decision making in the allocation of resources and responsibility. This is true at both the individual and the institutional level. This chapter describes both causes and consequences of these errors, including consequences already present in society. I make specific proposals for best incorporating robots into our society. The potential of robotics should be understood as the potential to extend our own abilities and to address our own goals.

In this chapter I focus on the ethics of building and using non-human artificial Companions. The primary topic of this book is digital Companions, not conventional robots, but both pragmatically and ethically the issues are the same. A robot is any artificial entity situated in the real world that transforms perception into action. If a digital assistant listens and talks to a human, it is a robot – it is an agent, an actor, living in and changing the world. My thesis is that robots should be built, marketed and considered legally as slaves, not Companion peers.

Digital agents not only change the world by affecting the people they converse with. They may also communicate what they learn to others – directly or indirectly through shared databases or others' agents. Agents transmit, create and may even destroy information, including human opinions and reputations. Digital agents may use the Internet to actively purchase goods or services, thus causing the movement of physical objects as well as ideas. Finally, some Companion agents really are conventional metal robots with legs and wheels. Such robots can do all the things a digital robot can do, and also produce direct physical impact on the world – from holding hands or washing windows to breaking dishes and falling down stairs. One aspect of direct physical impact is an increased sense

of presence for the humans. We are, after all, animals with evolved facilities for perception that automatically provide assessment and rewards for social actions.

The question I focus on in this chapter is, what is the correct metaphor we should use when thinking about our relationship with robot Companions? By turns in this article I will use two different definitions of *correct*. First, there is the question of what is *accurate* – what are we really producing? And second, what is *appropriate* – what should we try to build, and how should we try to sell it?

Robot-oriented ethics are fundamentally different from ethics involving other intelligent entities, because they are by definition artifacts of our own culture and intelligence. Perhaps unfortunately, we actually have almost as much control over other species and sometimes peoples as we do over robots. We as a culture do regularly decide how much and many resources (including space and time) we willingly allocate to others. But biological species have exquisitely complicated and unique minds and cultures. If these minds and cultures are eliminated, they would be impossible to fully replicate. In the case of robots, the minds are not there yet, and the culture they would affect if we choose to allow them to will be our own.

Why slaves?

Slaves are normally defined to be *people you own*. In recent centuries, due to the African trade, slavery came to be associated with racism and also with endemic cruelty. In the past though (and in some places still today) slaves were often members of the same race or even nation that had simply lost private status. This happened generally as an outcome of war, but sometimes as an outcome of poverty. Excesses of cruelty are greatest when actors are able to dehumanise those in their power, and thus remove their own empathy for their subordinates. Such behavior can be seen even within a small contemporary community of citizen peers, when a person in power considers their social standing as an indication of a specialness not shared with subordinates. Our culture has for good reason become extremely defensive against actions and the beliefs associated with such dehumanisation.

But surely dehumanisation is only wrong when it's applied to something that really is human? Given the very obviously human beings that have been labelled *inhuman* in the global culture's recent past, many seem to have grown wary of applying the label at all. For example, Dennett (1987) argues that we should allocate the rights of agency to anything that *appears* to be best reasoned about as acting in an intentional manner. Because the costs of making a mistake and trivializing a sentient being are too great, Dennett says we are safer to err on the side of caution.

Dennett's position is certainly easy to sympathize with, and not only because such generosity is almost definitionally nice. As I discuss below, there are many reasons people want to be able to build robots that they owe ethical obligation to. But the position overlooks the fact that there are also costs associated with allocating agency this way. I describe these costs below as well.

But first, returning to the question of definition – when I say "Robots should be slaves", I by no means mean "Robots should be people you own." What I mean to say is "Robots should be *servants* you own."

There are the fundamental claims in this paper:

1. Having servants is good and useful, provided no one is dehumanised.
2. A robot can be a servant without being a person.
3. It is right and natural for people to own robots.
4. It would be wrong to let people think that their robots are persons.

A correlated claim to the final one above is that it would also be wrong to build robots we owe personhood to. I will not discuss that point at length here; I have at least attempted to make that case before (Bryson, 2000). But this corollary follows naturally from my final claim above, so I will return to it briefly towards the conclusion of this chapter.

Why we get the metaphor wrong

There is in fact no question about whether we own robots. We design, manufacture, own and operate robots. They are entirely our responsibility. We determine their goals and behavior, either directly or indirectly through specifying their intelligence, or even more indirectly by specifying how they acquire their own intelligence. But at the end of every indirection lies the fact that there would be no robots on this planet if it weren't for deliberate human decisions to create them.

The principal question is whether robots should be considered strictly as servants: As objects subordinate to our own goals that are built with the intention of improving our lives. Others in this volume argue that artificial Companions should play roles more often reserved for a friend or peer. My argument is this: given the inevitability of our ownership of robots, neglecting that they are essentially in our service would be unhealthy and inefficient. More importantly, it invites inappropriate decisions such as misassignations of responsibility or misappropriations of resources.

Many researchers want to build AI that would have moral agency – that is, to which we would owe ethical obligations as we do to a person. Helmreich (1997) suggests this desire to create Artificial Life is most present in researchers that

are middle-aged males with a fixation on an individual capacity for creating life. While there seems to be a certain empirical plausibility of Helmreich's hypothesis, in my own experience many people who are not themselves roboticists or middle-aged still seem to have a shockingly strong tendency to believe they owe ethical obligation to robots.

I was astonished during my own experience of working on a (completely non-functional) humanoid robot in the mid 1990s, by how many well-educated colleagues volunteered – without prompting and immediately on seeing or even hearing about the robot – that unplugging such a robot would be unethical. Less anecdotally, popular culture contains many examples of heroic, conscious robots examining the worth of their own lives. For example, in the original *Star Wars (A New Hope)*, there is a running theme concerning the slavery and even torture of robots which is at times explicitly voiced by the 'droid character C-3PO. This theme is not resolved in that movie, nor is it so prominent in any of the *Star Wars* sequels. However in *Blade Runner*, *The Bicentennial Man*, *A.I.: Artificial Intelligence,* and several episodes of *Star Trek: The Next Generation* featuring the robot crew member Data, the central question is what it takes for a robot to be a moral agent. Traditional literary criticism of science fiction holds that the artistic examination of alien or artificial sentience is a mechanism for examining by proxy humanity and the human condition. However, many producers and consumers of science fiction consider themselves futurists, examining future rather than (or along with) present moral dilemmas. Whatever the artistic intention, from my experience as a roboticist, it seems that a large proportion of science fiction consumers are comfortable with a conclusion that anything that perceives, communicates and remembers is owed ethical obligation. In fact, many people seem not only comfortable, but proud of having come to this realization, and willing to defend their perspective angrily.

Bryson and Kime (1998) have argued that both deep ethical concern for robots, and unreasonable fear of them, results from uncertainty about human identity. Our identity confusion results in somewhat arbitrary assignments of empathy. For example, contemporary cultures – including some scientific ones – often consider language both necessary and sufficient for human-like intelligence. Further, the term *conscious* (by which we mostly seem to mean "mental state accessible to verbal report") is heavily confounded with the term *soul*, (meaning roughly "the aspect of an entity deserving ethical concern"). Thus, for example, animal 'rights' debates often focus on whether animals are conscious, with the explicit assumption that consciousness automatically implies ethical obligation (Rollin, 1998). Of course, neither the term *consciousness* nor *soul* is precisely or universally defined. Rather, these terms label concepts formed by accretion over millenia as we have attempted to reason about, analyse and describe ourselves.

I would see Helmreich's diagnosis then as just part of a much wider problem. It is not only that a few prominent middle-aged artificial life researchers have fantasies of autonomously creating human-like life. Entire cultures – including perhaps the global one – could be expected to take pride in such an achievement, as they did in the conquest of space or in colonialism and empire. But this desire is a symptom of a larger confusion about whether AI could or should be a surrogate for human life and human responsibility.

Since this book examines artificial Companions, in the rest of this chapter I concentrate primarily on the best role for robots in the household – the costs and benefits of different attitudes towards robots at the individual and commercial level.

Costs and benefits of mis-identification with AI

The individual level

Over-identification with AI leads to a large range of category errors which can significantly bias decision making. This is true at decision-making levels ranging from the individual to the national and super-national. Already we have seen commercial willingness to exploit human empathy for AI objects such as Tamagotchi 'pets'. Like every other potential distraction – from radios to children – Tamagotchi have led to fatalities from automobile accidents. But there are more important policy considerations than the occasional sensational headline.

At the personal level, the cost of over-identification with AI should be measured in:

1. The absolute amount of time and other resources an individual will allocate to a virtual Companion,
2. What other endeavors that individual sacrifices to make that allocation, and
3. Whether the tradeoff in benefits the individual derives from their engagement with the AI outweigh the costs or benefits to both that individual and anyone else who might have been affected by the neglected alternative endeavors.

This final point about the tradeoff is somewhat convoluted, so I want to clarify why I make so many qualifications. My arguments in this chapter derive primarily from the default liberal-progressive belief that the time and attention of any human being is a precious commodity that should not be wasted on something of no consequence. However, life is more complicated than simple principles. Clearly some people are socially isolated, and it has long been demonstrated that isolated elderly people really are healthier with empathic, social non-human Companions

(Siegel, 1990). Also well in evidence is that some people sometimes engage in harmful, antisocial behavior. There is at least preliminary evidence that increased Internet access directly correlates to the recent impressive decreases in the levels of sexual assault (D'Amato, 2006). Thus it is at least possible that in some cases the 'cost' of over-identification with AI could in fact be negative. In other words, becoming overly emotionally engaged with a robot may in some cases be beneficial, both for the individual and society.

But returning to the default liberal-progressive view, my concern is that humans have only a finite amount of time and drive for forming social relationships (Dunbar, 1997), and that, increasingly, we find ways to satiate this drive with non-productive faux-social entertainment. Putnam (2000) documents the massive decline in what he calls "social capital" in the Twentieth Century. Although Putnam originally ascribed this decline to increasingly dynamic societies and a lack of 'bridging' individuals between communities, his own further research failed to sustain this theory. A simpler explanation seems likely: that the drive to socialize is increasingly being expended on lower-risk, faux-social activities such as radio, television and interactive computer games. Each of these technologies has progressively increased the similarity between individual entertainment and true interpersonal interactions.

More recently online social networking has reintroduced a human element to technological entertainment. Facilitating human-human interactions could be a component of a domestic robot's behavioral repertoire, too. However, this complication is irrelevant to the main argument of the present chapter. I am considering here autonomous robots – actors in their own right. Virtual or physical avatars being teleoperated by their owners' friends or family are already slaves in such an extreme sense that I don't believe anyone would argue they deserve moral agency. Such robots, called *avatars* performs precisely the actions determined for them by the people they represent, and have no internal motivation whatsoever.

The institutional level

The individual-level social cost of mis-identification with robots is the economic and human consequence of time, money and possibly other finite resources being given to a robot that would otherwise be spent directly on humans and human interaction. The cultural – or national – level costs obviously include the combined individual costs (or benefits) of the citizens of that culture or nation. But it is possible that in addition similar errors might also arise at a higher institutional level. I am not particularly concerned that the enthusiasm for the creation of life, or of super-human, super-obedient robot citizens, might actu-

ally attract government research funding or legislative support. However, both the USA and the European Union have openly made research on autonomous cognitive systems a high funding priority. In the US this has been accompanied by both funding and media attention on making certain the systems (in the US, mostly battlefield robotics) are capable of moral decision making. Government-funded robotics professors are openly suggesting that robots might make more ethical decisions than humans in some battlefield situations (Wallach and Allen, 2008; Fong et al., 2008).

While I have no problem with the use of Artificial Intelligence to complement and improve human decision making, suggesting that the AI itself *makes* the decision is a problem. Legal and moral responsibility for a robot's actions should be no different than they are for any other AI system, and these are the same as for any other tool. Ordinarily, damage caused by a tool is the fault of an operator, and benefit from it is to the operator's credit. If malfunctions are due to poor manufacturing, then the fault may lie with the company that built it, and the operator can sue to resolve this. But creating a legal or even public-relations framework in which a robot can be blamed is like blaming the private soldiers at Abu Ghraib for being "bad apples". Yes, some apples are worse than others and perhaps culpably so, but ultimate responsibility lies within the command chain that created the environment those privates operated in. Where the subject is machines we build and own, then the responsible role of the organisation is even clearer. We should never be talking about machines taking ethical decisions, but rather whether machines operated correctly within the limits we set for them.

I can see no technological reason why people should expect moral agency from an autonomous mobile gun when they do not expect it from automatic tellers (banking machines) or conventional automatic dishwashers. Of course shooting is more dangerous than cleaning, but that doesn't make guns more moral agents than sponges. As we increase the on-board sensing, action and logic in these tools, can there really come a point where we may ourselves abrogate ethical responsibility for directing the taking of lives? In my opinion, no.

Advanced weapon systems may seem exceptional, but I expect these technologies and issues will rapidly move into the civilian sector. This was the history of air flight: Originally the domain of engineers and hobbyists, once flight was conquered it was immediately applied to warfare. Following World War I the technology was quickly exploited by the creation of an industry for both business and leisure services. If we allow robots to bear their own responsibility for their behavior in foreign battlefields, we will soon face the same issues at home as we find them working in police forces and post offices.

We do not have to wait for the presence of advanced AI to see the consequences of responsibility passing away from humans. Consider existing automated (or

at least unstaffed) railway stations. Ordinarily they work well, perhaps providing ticket services in more languages than a small rural station might otherwise provide. But when a train fails to make a scheduled stop at six on a Sunday morning, there is no one to apologise and provide a replacement taxi service, making sure you get to the airport for your flight.

At the national and institutional level, this is the moral hazard of being too generous with personhood. Not only does automation save in staff cost, but also it reduces corporate responsibility for service from the level of the reasonable capacity of a human to that of a machine. Further, the capacities of a machine are largely determined by those who choose how much they will pay for it. At the personal level, the moral hazard is choosing less rich interactions with lower social impact because robotic interactions are more predictable and less risky. If you do something stupid in front of a robot, you can delete the event from its memory. If you are tired of a robot or just want to go to bed early, you don't need to ask its opinion, you can just turn it off.

While there may well be people and institutions for whom such automated services are the only available option, the question is how many will be tempted unnecessarily into being less responsible and productive members of society. As with the level of government welfare benefits, there probably is no ideal target level of faux-humanism we can provide in a robot that both helps everyone who really needs assistance, while not tempting anyone away from social contribution who does not. Finding the appropriate level then is difficult, but my claim is that at least in the case of robotics, it can be located more easily by providing accurate public information.

Getting the metaphor right

Why, if robots are so hazardous should we want to include them in our lives? Because servants have the potential to be useful. When there was greater income disparity in the UK and domestic chores such as cooking were far more time consuming, human servants were much more common. Laslett (1969) found that approximately 30% of all households had servants in a set of British villages surveyed from 1574–1821. At that time, tasks such as cooking and cleaning for even a moderately-sized family could take two people 12 hours a day every day (Blair, 2008). Where wives and other kin were not available to devote their full time to these tasks, outside employees were essential.

Contemporary cooking and cleaning takes far less time than in the days before electricity, gas or running water. Nevertheless, many aspects of food preparation are still often outsourced. Now this work is done in factories, by

agricultural labour or in restaurants rather than by in-home domestic servants. The European Commission has chosen to invest heavily in cognitive systems for domestic robotics, partly due to anticipated foreign demand and perceived local expertise, but also in an effort to increase the productivity of its own workforce.

Anyone who has worked recently with robots knows that they are unlikely to be deployed dusting plants or cleaning fine china in the near future. The most likely initial introduction of domestic robotics will be partial or total replacement of expensive human help in tasks that require less real-time planning and dexterity. Examples are physical support for the infirm, minders and tutors for children, and personal assistants for those with challenged working memory capacity, whether that challenge is due to disease or to distraction. Note that none of these applications requires anthropoid robots. In fact, a personal assistant might take the form of a personality-based video interface for a smart home. Avatars for the AI might flicker to the nearest fixed screen location when called; AI may monitor embedded household sensors for danger, injury, or just a forgotten lunch. In homes with more than one occupant, a variety of robot personalities can use the same infrastructure, each serving their own user.

In my opinion, communicating the model of robot-as-slave is the best way both to get full utility from these devices and to avoid the moral hazards mentioned in the previous sections. Of course some people will still talk to their robots – some people talk to their plants and others their door knobs. But those people have neighbours and relatives who know that the plants and door knobs don't understand. These people help their overly-conversant relative or friend keep that fact somewhere in their minds. Similarly, our task is not to stop people from naming or petting their robots. Our task is to ensure that the vast majority of the population understands that robots are just machines, and that one should spend money and time on them as is appropriate to their utility, but not much more.

Of course, some people may for a variety of personal, historical or cultural reasons object to having any form of servant or slave in their home. In that case, there is another metaphor that might be useful – that of the extended mind (Clark and Chalmers, 1998). If the robot has no goals except for those it assumes from you, then there are rational arguments to be made that the robot is just an extension of yourself. I have previously made an analogous argument that the Semantic Web should be viewed neither as a giant database nor as a set of agents to be knitted together in a complex democratic community in order to provide a simple service. Rather, the Semantic Web should be thought of as snippets of intelligence that can be used to augment the capabilities of a user's own personal assistant (Bryson et al., 2003). Only the personal assistant has a motivational structure, and this it inherits from the goals of its user.

It is a fairly simple extension of this argument to say that the digital personal assistant itself could be considered an extension of the user. Our goals, beliefs, perception and capacity for action can all be extended or made more reliable through a range of robotic servants – or if one prefers, services. In extending what we do and know, robots could also be seen to extend who we are.

Don't we owe robots anything?

If servants are such a great idea, shouldn't we just hire more human ones? To some extent, we already do. Our economy has a large service component. Many tasks that a hundred years ago were performed by live-in servants are now largely performed by strangers outside the home, such as food preparation and clothes manufacture. Other tasks are already performed at least in a large part by machines, such as washing laundry, mowing lawn, or keeping an appointment calendar.

The poor are richer now than they once were. Virginia Woolf paid her live-in servants only 1% of her own annual income of £4,000 (Blair, 2008). Even a part-time domestic servant who was willing to take 1% of a modern professional salary would be unlikely to do many things that could not better be done with machines or outside services.

But the most difficult thing with human servants is of course the fact that they really are humans, with their own goals, desires, interests and expectations which they deserve to be able to pursue. Humans living and working together but set as not each other's equals are often vulnerable to frustration and exploitation.

But what about the robots? Would they not feel frustrated? Would they not be exploited and abused?

Remember, robots are wholly owned and designed by us. We determine their goals and desires. A robot cannot be frustrated unless it is given goals that cannot be met, and it cannot *mind* being frustrated unless we program it to perceive frustration as distressing, rather than as an indication of a planning puzzle. A robot can be abused just as a car, piano or couch can be abused – it can be damaged in a wasteful way. But again, there's no particular reason it should be programmed to mind such treatment. It might be sensible to program robots to detect and report such ill treatment, and possibly to even avoid its abuser until its owner has been notified (assuming the abuser is not the owner). But there is no reason to make a robot experience suffering as part of the program to generate such behavior.

Owners should not have ethical obligations to robots that are their sole property beyond those that society defines as common sense and decency, and would apply to any artifact. We do not particularly approve of people destroying rare,

fine cars with sledge hammers, but there is no law against such behavior. If a robot also happened to be a particularly fine piece of art then we would owe it the same obligations we owe other pieces of art.

Robot owners should not have obligations, but ensuring this is the responsibility of robot builders. Robot builders *are* ethically obliged – obliged to make robots that robot owners have no ethical obligations to (Bryson, 2000). A robot's brain should be backed up continuously offsite by wireless network; its body should be mass produced and easily interchangeable. No one should ever need to hesitate an instant in deciding whether to save a human or a robot from a burning building. The robot should be utterly replaceable. Further, robot owners should *know* their robots do not suffer, and will never 'die' even if the rest of their owner's possessions are destroyed. The robot's brain state should be preserved off site. The robot can return to function exactly as before as soon as a new body can be acquired, though it may need some retraining if there is a new domicile to inhabit or slight variations between bodies. Robots then can be a relied on as no more than extensions of their owners. They should not be anthropoid if that can be helped, and their owners should have access to the robots' program-level interface as well as its more socially-oriented one. This will help the owners form a more accurate, less human model for reasoning about their Companions.

We do then have obligations regarding robots, but not really to them. Robots are tools, and like any other artifact when it comes to the domain of ethics. We can use these tools to extend our abilities and increase our efficiency in a way analogous to the way that a large proportion of professional society historically used to extend their own abilities, but with fewer ethical and logistical hazards. Hopefully, we can continually increase the number of service owners in our society, so a smaller proportion of everyone's time can be spent on mundane or repetitive tasks if they do not enjoy them. In that case, a larger proportion of time and resources can be spent on useful processes, including socializing with our colleagues, family and neighbours.

Conclusions

Why do people want robots to be their peers? Is it perhaps because they want a 'peer' that will never argue, or at least never be smug when it wins? A fairy godparent smarter than themselves, that they can nevertheless ultimately boss around and pen up like a pet dog? If so, such narcissism is probably mostly harmless, and perhaps a good thing for the dogs. But in a liberal democracy we tend to think of every citizen's life and mind as a valuable resource. Wasting that resource 'socializing' with artifacts would be a great loss.

Robots should rather be viewed as tools we use to extend our own abilities and accelerate progress on our own goals. An autonomous robot definitionally incorporates its own internal motivational structure and decision mechanisms, but *we* choose those motivations and design the decision-making system. All their goals are derived from us.

I have argued here that robots are often overly personified. First, this is because of our desire to have the power of creating life. Second, this is because we are not certain what it means to be human, so we currently offer the term to anything that senses, acts, remembers and speaks. Given the errors in dehumanisation that have been broadly made in the very recent past – in fact, sometimes in the present – the desire to avoid such mistakes is laudable. Yet ironically, extending the title *human* to something which is not only serves to further devalue real humanity.

The objective of this chapter has been to persuade roboticists and robotophiles – now, while this industry is in its early stages – that calling a robot a moral agent is not only false but an abrogation of our own responsibility. I have also demonstrated that these problems are already in our society: in the current research funding for ethical battlefield robots, and in the commercial exploitation of human empathy for artificial characters. My conclusion is that we are obliged not to the robots, but to our society. We are obliged to educate consumers and producers alike to their real obligations with respect to robotics.

Acknowledgements

This chapter was written while on sabbatical from the University of Bath on a fellowship from the Konrad Lorenz Institute for Evolution and Cognition Research in Altenberg, Austria. Thanks to both institutions for the time to work on this project.