The Rise of the Robots and the Crisis of Moral Patiency

By John Danaher, NUI Galway

(Pre-publication version of *AI and Society* DOI: 10.1007/s00146-017-0773-9)

This paper adds another argument to the rising tide of panic about robots and AI. The argument is intended to have broad civilization-level significance, but to involve less fanciful speculation about the likely future intelligence of machines than is common among many AI-doomsayers. The argument claims that the rise of the robots will create a crisis of moral patiency. That is to say, it will reduce the ability and willingness of humans to act in the world as responsible moral agents, and thereby reduce them to moral patients. Since that ability and willingness is central to the value system in modern liberal democratic states, the crisis of moral patiency has a broad civilization-level significance: it threatens something that is foundational to and presupposed in much contemporary moral and political discourse. I defend this argument in three parts. I start with a brief analysis of an analogous argument made (or implied) in pop culture. Though those arguments turn out to be hyperbolic and satirical, they do prove instructive as they illustrates a way in which the rise of robots could impact upon civilization, even when the robots themselves are neither malicious nor powerful enough to bring about our doom. I then introduce the argument from the crisis of moral patiency, defend its main premises and address objections.

Keywords: Robotics; Artificial Intelligence; Technological Unemployment; Algocracy; Moral Agents; Moral Patients

1. Introduction

Look around: the robots are on the rise and they are not the slow, lumbering giants of science fiction. They are smart, agile and flexible, taking advantage of developments in machine learning and artificial intelligence. They take on a variety of forms, some humanoid, some not. They are embedded into more and more aspects of our everyday life (Ford 2015; Dormehl 2014; Carr 2015).

Does this spell disaster for our civilization? Some people are worried. At the more extreme end of the spectrum, there are those that believe that the rise of intelligent machines could pose an existential threat to humanity (Bostrom 2014). A smarter than human machine, it is argued, could have a decisive strategic and practical advantage over humanity and could control the future of life on this planet in a value-altering way. At the slightly less extreme end of the spectrum, there are those that worry about the effect on employment and joblessness (Frey and Osborne 2013; Brynjolfsson and McAfee 2014; Ford 2015; Avent 2016). In between, there are a variety of more specific and prosaic concerns. Some are focused on the impact of robotics on the law (Calo, Froomkin and Kerr 2016); and on the ethics of warfare (Bhuta et al 2016); and on our ability to think for ourselves (Carr 2015). Others are concerned with how robots and AI will change the nature of our social and intimate relationships; and about the moral agency and responsibility of robots themselves.

This article adds another argument to the rising tide of panic, albeit one that I claim is both more subtle and disturbing than others in this debate. The argument is intended to have broad civilization-level significance, but to involve less fanciful speculation about the likely future intelligence or power of intelligent machines than is common among many AI-doomsayers. The argument claims that the rise of the robots could lead to a *crisis of moral patiency*. That is to say, it could compromise both the *ability* and *willingness* of humans to act in the world as responsible moral agents, and consequently could reduce them to moral patients. Since the ability and willingness to act as a responsible agent is central to the value system in modern liberal democratic states, the crisis of moral patiency has a broad civilization-level significance: it threatens something that is foundational to and presupposed in much contemporary moral and political discourse. In defending this argument, I use the terms ‘patiency’ and ‘agency’ in slightly non-standard ways. Most philosophers view them as statuses that an

entity either possesses or not and not as properties that can wax or wane. I prefer the latter view and I defend this non-standard interpretation as part of my general case.

The argument comes in three parts. I start with a brief analysis of analogous arguments that have been made (or implied) in popular culture, specifically in the TV series *Futurama* and the movie *Wall E*. I do this to shift the reader's perspective at the outset. Though the arguments that derive from these pop culture sources are hyperbolic and satirical, they do get us to think about ways in which the rise of robots could impact upon civilization, even when the robots themselves are neither malicious nor powerful enough to bring about our doom. I then introduce my own argument — the argument to the crisis of moral patiency — and defend its main premises from a variety of objections. Finally, I conclude by assessing the implications of the argument for the future.

Before I begin, I should offer some clarification of the terminology used in this article. Throughout, I refer, interchangeably, to the rise of the 'robots' and the rise of 'AI'. This might confuse some people insofar as robots and AI are separable (to at least some extent). Robots are physical objects that are able to act in the world. They usually have some degree of artificial intelligence, but AI itself need not have the power to act in the real world. Indeed, if anything, the norm is against this. AI is typically confined to a computer, and often just assists or offers instructions to human actors. I refer to both phenomena interchangeably in this article because my argument covers both the robot with powers to act in the world and the AI confined to a box issuing instructions and recommendations. In other words, I refer to both because I think the crisis of moral patiency derives from our increasing reliance on both technologies.[1]

Another point of clarification concerns my use of somewhat emotive and hyperbolic language: the *rise* of the robots; the *crisis* of moral patiency. My use of such phrases may put some people off, and cause them to discount what I have to say. To some extent, that reaction is fair: I am using that language to provoke and stimulate debate. But this should not detract from the seriousness of my argument. I am not using the language in a 'tongue-in-cheek' manner, or in the interests of satire. I believe that there is a serious risk to contend with here. One we would be well-advised to address.

[1] For a useful taxonomy of different possible interactions between humans and robots and AI, see Van de Voort, Pieters and Consoli (2015).

My seriousness should become clear when my argument is contrasted with arguments that are deliberately satirical.

2. *Futurama, Wall E* and the Collapse of Civilisation

I start by looking at two such satirical arguments. The first is taken from the TV series *Futurama*; the second from the movie *Wall E*. Both point to similar consequences from the widespread use of robots.

First, let's consider the argument from *Futurama*. Robots can and do take many forms. One of the most popular forms — among futurists, technologists and the wider public — is that of the sexbot. That is, a fully humanoid artificial being designed to be used for sexual pleasure.[2] In its third season episode 'I Dated a Robot', the TV show *Futurama* explores the possible civilization-ending effects of this technology. The show, as a whole, tells the story of a man named Fry who is cryogenically frozen on midnight 1999 and resuscitated in the year 3000. Most episodes of the series play upon his 'fish out of water' status. The episode 'I Dated a Robot' is no different in this regard. In it, Fry acquires a robot in the image of the actress Lucy Liu. He then starts dating her. His friends start to worry about the relationship, fearing that it is causing Fry to shut himself off from the world. They also know something he doesn't know: they know the serious threat posed by romantic (sexual) relationships with robots.

To educate Fry about the dangers, they show him a video entitled 'Don't Date Robots!'. The video is modeled on the classic style of an American public service announcement. In the video, we see an ordinary American teenager (Billy) who downloads a 'Monroe'-bot (a robot that looks like Marilyn Monroe). He starts 'making out' with this robot in his bedroom. His parents come in and ask him if he would like to clean his room, or get a job and earn some money. He responds that he would rather just 'make out with [his] Monroe-bot'. Then the girl from across the street comes in and asks if he would like to come across to her house later that night to 'make out'. He again responds in the negative, saying that this is an awfully long way to go just to 'make out'. The broadcast then makes its key argument. Because he has ready access to a romantic and sexual partner (in the shape of his Monroe-bot), Billy has lost the

[2] It is a long-standing trope in science fiction. Pris (the "pleasure model") featured in the film *Blade Runner*. And recent TV series such as *Westworld* and *Humans* have also featured robots used for sexual reasons.

motivation to do anything of real consequence. He can just sit in his room and get all the pleasure he wants from his life. This spells disaster for civilization. As the video then makes clear, everything that is valuable about human civilization (and some of what is not so valuable) is down to the struggle to secure a romantic sexual partner. This includes art, music, science, technology and sport. All the things we like to celebrate as hallmarks of our civilised existence. They are wiped out by the ready availability of sex robots.

In more formal terms, the argument from this episode of Futurama seems to be the following:

> (1) The fact that it is difficult to secure a romantic and sexual partner is what sustains everything we value about society (art, music, science, technology, sport etc).
>
> (2) If sex robots become viable and attractive alternatives to human romantic and sexual partners, and if they become widely and readily available, they will remove the difficulty of securing a romantic and sexual partner.
>
> (3) Therefore, if sex robots become viable and attractive alternatives to human romantic and sexual partners, and if they become widely and readily available, everything we value about civilization will collapse.

Amusing and fascinating though it is, the argument is clearly *hyperbolic* and *unsound*. Premise (1) has some tincture of credibility: it probably is true to say that the desire for a romantic and sexual partner, and the difficulty of securing such a partner, motivates *some* of the activities that we value, but it is unlikely to be the *sole* motivation (contra Freud). Art and science might be their own reward, and the desire for other social goods like power and influence is likely to motivate as well. Added to this, there are problems with premise (2). It contains the caveat that the robots must be 'viable and attractive alternatives' to human partners, but achieving this in the real world may be difficult. It is possible that robot sex will never be a simple substitute for human sex. There may be psychological and technical barriers to making this a reality.[3] It also contains the caveat that the robots must be widely and readily available. This would appear to be necessary

[3] Some of these, including the so-called uncanny valley, are discussed in Danaher (2014).

if the pan-civilizational impacts are going to materialise, but this too could prove difficult. The technology needed to create robotic partners may be expensive and not available to all.

Nevertheless, and despite its problems, the Futurama argument does say something interesting about the rise of the robots. It suggests that the rise of the robots could lead to a decline in humans' willingness to express their moral agency (to make significant moral changes to the world around them). Because they have ready access to pleasure-providing robots, humans might become increasingly passive recipients of the benefits that technology bestows. It also highlights how this decline in agency may cut us off from some of the most valuable aspects of our modern civilization. And, critically, it does all this without assuming that the robots are all-powerful, malicious, terminator-like beings. In this way, it highlights a civilization-level threat that does not materialize from superintelligent robots. In essence it is saying that the robots themselves may not be problem; rather, the subtle way in which they play upon our psychological biases and temptations may be the problem.

The Pixar movie *Wall E* suggests something very similar, only it has nothing to do with sex or the drive for procreation. The movie is largely silent: there is little dialogue. It doesn't try to make any explicit arguments or claims like the PSA in *Futurama*. But there are clearly arguments embedded within its narrative. Indeed the movie contains reasonably overt ecological and technological critiques of humanity. It depicts a future in which the Planet Earth has been completely despoiled and turned into a metallic junkyard. The human population has escaped onboard large, interstellar spacecraft, bound for a new home. Onboard these spaceships, practically useful, artificially intelligent robots are everywhere. They fly the ship and maintain its operations. The humans are mere passengers in every sense of the word. They have become obese, slug-like entities, carried about the ship on motorized chairs, constantly fed an endless stream of junkfood and entertainment by the technological infrastructure they have created. They are passive recipients of the benefits of technology, not active agents changing the world in which they live. The argument implicit in this depiction seems to be that if robots can perform all the tasks that humans used to perform – and if they can do so in a more cost-efficient and effective manner – there will be little incentive left for humans to do anything except sit back and enjoy the ride. Their willingness to act as agents is likely to be diminished by the technological abundance.

So, once again, we see a fictional representation of a decline of agency being wrought by the robotics revolution. And, once again, this dramatic effect is being wrought by technology that is largely benign and well-intentioned, i.e. the robots in *Wall E* seem to genuinely care for their human passengers.

But, of course, the argument implicit in the movie *Wall E* is intended as satire, not as something that is necessarily philosophically or empirically credible. What if there were a more plausible argument that defended similar claims? That's what the argument from moral patiency seeks to provide.

3. The Argument to the Crisis of Moral Patiency

To understand the argument, we need to clarify the distinction between a moral agent and a moral patient.[4] The concept of a moral agent is relatively familiar. An agent is any being that can acquire information from its surroundings, can process that information and use it to develop goals and action plans, and has the ability to act in a manner that implements those action plans. A moral agent is any being who is capable of taking moral ownership or responsibility for those actions. Exactly what is required for this moral ownership is contentious. It is pretty clear that some capacity for moral understanding (i.e. understanding of moral rules and values) is essential. It is also pretty clear that some ability to regulate one's actions in light of moral rules and values is essential.[5] But beyond that there is much to debate. Some people argue that freedom of the will is essential, some that consciousness is a pre-requisite, others that intention-guided action is enough. We need not engage those complexities here. It is enough for us to presuppose that human beings are capable of being moral agents.[6] As I will explain in a moment, this is important because moral agency is foundational to the value-structure of modern liberal-democratic states.

[4] Perhaps the best discussion of the distinction, with a particular focus on robotics, is to be found in Gunkel (2012); see also Floridi (1999).

[5] The previous two sentences articulate the classic Aristotelian account of moral agency which is based on the satisfaction of two conditions: (i) a control condition and (ii) a knowledge condition.

[6] I avoid saying that they *are* moral agents on the grounds that I take moral agency to be akin to a capacity that one has the power to exercise, but which one may not always exercise. This is important for the argument I wish to make. If humans simply are moral agents, and they never lose this status, then nothing about the rise of robots can take that away from them. But if their moral agency is like a muscle that must be exercise lest it atrophy, the argument is wish to make can work.

The concept of a moral patient is perhaps less familiar. As Gunkel (2011, 94) and Hajdin (1980, 180) note, it is usually defined by way of contrast to that of the moral agent. A moral patient is thus a being who possesses some moral status — i.e. is owed moral duties and obligations, and is capable of suffering moral harms and experiencing moral benefits — but who does not take ownership over the moral content of its own existence. To put it another way, a moral patient is a being upon whom well-being (and other valuable states) are bestowed, but which does not (or cannot) take an interest in the autonomous formulation and pursuit of its own moral goals.[7] There is much debate about who or what is a moral patient. It is generally accepted that humans, in virtue of their moral agency, are also moral patients (the latter being a component of the former); it is also widely accepted that animals (or at least certain animals) are moral patients in virtue of their capacity to experience pleasure and pain. The patiency (and agency) of robots is more vigorously contested at present (Gunkel 2011). But this debate is irrelevant to the present argument. All that matters to me is that humans are capable of being moral patients. Although the interaction between humans and robots is relevant to my inquiry, the actual moral status of the robots is not.

The argument I wish to make is that the rise of the robots will lead to the decline of human moral agency and the rise of moral patiency. This might seem a little odd in light of the definitions just offered and the standard interpretation of agency and patiency in the philosophical literature. Surely, the critic will argue, agency and patiency are statuses that beings either possess or they don't, they are not qualities that can increase or decrease in magnitude? In one sense, this is correct. I accept that humans *are* moral agents and that they *are* moral patients, and that they likely will always be moral agents and moral patients. The widespread deployment of robots cannot take those statuses away. However, I argue that even though this is true, it makes sense to conceive of agency and patiency as statuses that can wax and wane. Sometimes the agency-like qualities are in the ascendency and sometimes the patiency-like qualities. It is in virtue of certain capacities (understanding, intelligence, rationality, the power to act in the world) that humans count as agents and it is virtue of certain other capacities (to feel pain, to have their interests thwarted) that they count as patients. Those capacities can be impaired or limited in certain environments or at certain stages of maturation. Children and those with severe disabilities, for instance, exhibit more of the characteristics of moral patients than moral agents: they have less understanding and

[7] This is a paraphrase of Peter, F. (2008) at 36.

ability to act in the world, but nonetheless can suffer pain and have their interests thwarted. Consequently, I argue that it makes sense to say that an individual is more or less patient/agent-like. What's more, sometimes the individual human being is more motivated or willing to take interest in exercising their power of moral agency: they take great care to learn the moral rules and gain moral insight; they carefully plan and implement actions that will make the world (or their lives) better. Think of the young philosophy student who goes out and reads Peter Singer's latest defence of *effective altruism*, critically scrutinises the arguments, persuades herself that they are correct, and then makes it her life's mission to ensure that she does the most good through her charitable giving. They are displaying their agency-relevant capacities to the highest degree. Contrariwise, sometimes humans are motivated to express more of the patiency-like qualities and the agency is minimised. The individual human becomes a largely passive recipient of the moral beneficence of others. Again, early stages of childhood are like this, as are certain states of illness and disability.[8]

My claim then is not that the rise of the robots will cause humans to lose the status of moral agency and acquire the status of moral patiency. Instead, my claim is that the rise of the robots will push the agency-like qualities to the background by reducing the need and desire for their expression.[9] This will lead to the ascendency of the patiency-like qualities, and will threaten many of the underlying values of our civilization. To put it in simple, formal terms:

> (4) Anything that threatens to suppress our moral agency (and increase the expression of moral patiency), threatens some of the foundational values of our civilization.
>
> (5) There is good reason to think that current and future trends in robotics and AI will suppress our moral agency and increase the expression of moral patiency.

[8] Not all states of disability and illness undermine agency. Recognition of this fact is extremely important in light of the negative history of the treatment of those with disabilities.

[9] Another criticism one could offer here is that framing my argument in terms of agency and patiency is misleading since there is already a perfectly serviceable vocabulary for expressing the argument, namely the vocabulary of activity and passivity. Robots render us more passive beings and less active ones. It is true that my argument does concern itself with activity and passivity, but this vocabulary fails to do justice to what I am arguing because it fails to identify the linke between specific forms of passivity and moral agency/patiency and then fails to spot the link between agency/patiency and foundational civilizational values. The vocabulary of patiency and agency draws out this important link.

(6) Therefore, there is good reason to think that current and future trends in robotics and AI threaten the foundational values of our civilization.

My main interest lies in the defence of premise (5), and with illustrating why the suppression of moral agency is a plausible and realistic extrapolation from current developments in robotics and AI. But premise (4) is also clearly crucial. What reason is there for believing that agency is so central to our civilisation's value structure? The simple answer is that the concept has always featured heavily in Western moral and political philosophy. Three examples illustrate its importance. First, as Floridi (1999) notes, the virtue ethical tradition that originates with the Ancient Greeks is essentially an agency-based theory of the good. It contends that the good life consists in developing agency-relevant traits like courage, honesty, and temperance. If we accept this tradition, then the decline of agency impairs our ability to live the good life. Second, agency is central to the moral ethos of the liberal democratic state. It is the pre-supposed capacity for moral agency that renders coercive interference with the citizens of a liberal state impermissible; it is the exercise of this capacity which makes political organisation possible (and permissible) according to contractarian theories of political morality; and it is the sensitivity to moral agency that makes liberty-oriented rights so important in such states. If agency declines, then so too will the seeming importance of these rights. Finally, as Gunkel notes (2011, 89), the history of moral progress has largely been the history of the struggle to recognise moral agency in the 'other'. The two most prominent examples of this being the recognition of the agency of slaves and women. Thus, if the expression of agency were to decline, we would, in some sense, be undermining the moral progress for which others have fought so hard and which are so integral to our civilization's self-conception. These three examples illustrate that moral agency really is foundational to our current way of life. It is a quality and good that we have traditionally sought to recognize and promote. If it were suppressed, the moral fabric of our civilization would be undermined.

This brings us to premise (5). To understand this premise we need first to have a general sense of what the 'rise' of robots and AI involves. As touched upon above, standard conceptions of human *intelligence* and *agency* argue that these qualities are made possible by the fact that humans have four specific capacities: (i) the capacity for *sensing*, i.e. acquiring information from the world around us; (ii) the capacity for *processing*, i.e. categorising, sorting and rendering useful that information; (iii) the

capacity for *acting*, i.e. the ability to use the processed information to form and implement action plans; and (iv) the capacity for *learning*, i.e. the ability to grow and develop the other three capacities. Moral agency is simply the application of these capacities to moral knowledge and action.

The first key insight into the rise of the robots and AI is that these technologies are designed to either supplement or take over from one or more of these capacities. Thus, in an obvious sense, any growth or development in those technologies will impact upon, and possibly suppress, the foundations of human agency, at least in certain domains. Consider the current project to develop the self-driving car. At the moment, driving is an activity to which human agency is central. We are the ones that acquire data from the world, organise it into a sensible form, and use it to drive the car in the direction we wish to go. Classically, the car was simply an actuating device, i.e. a way of amplifying or enhancing our ability to implement our action plans. In this sense, it was an agency-enhancing device: it enabled us to do more than our natural bodies would have allowed. Nowadays, the car is a more sophisticated machine. It still has its agency-enhancing powers, but it also now reduces the need for the expression of certain agency-qualities. For example, onboard computers and sensors often assist us in acquiring information from the world and developing action plans (e.g. collision alarms). There are also mechanisms that allow the car to take some of the burden associated with the act of driving (e.g. cruise control; brake assist). The goal of the tech industry — if the likes of Google, Apple, Tesla, and Uber are to be taken seriously — is to eventually fully automate the process of driving. The head of the Google project once said that they did not want to have steering wheels in their self-driving cars because human drivers will be less safe (Griggs 2014). This suggests that the ultimate goal is to eliminate human agents from the realm of driving. The result of this will be a suppression human agency (in the specific domain of driving), not an amplification of it.

If the quest to fully automate driving were an isolated example, it would not be too problematic. True, driving is one domain in which humans must exercise their moral agency (by showing concern for fellow road users, driving safely, making their lives more enjoyable and so forth), but it is just one domain. Human intelligence and agency can be directed toward many things. If agency is lost in one domain, it can be discovered (or recovered) in another. Indeed, losing agency in one realm could have a

positive overall effect on agency in other realms. The civilization-level effect that I allude to in my argument will only be felt if there is a much broader suppression of agency, one that threatens all the major outlets for the expression of human moral agency. The problem is that plausible extrapolations from existing trends in robotics and AI could easily lead to a broader suppression of agency. In fact, I believe that there are three such trends that are likely, when combined, to have such an impact.

The first is the general trend towards technological unemployment. This is long-term structural unemployment that is directly attributable to advances in robotics and AI. Fears about the replacement of human labour by machine labour are as old as the industrial revolution (possibly older), but they have taken on a more persuasive guise in recent years. A spate of publications have defended the impending reality of robotic takeover (Frey and Osborne 2013; Brynjolfsson and McAfee 2014; Ford 2015; Avent 2016). Historically, such fears have been addressed with the response that we will find alternative forms of employment. And this more optimistic view has been proven correct in the past. But there are reasons for thinking that 'this time it's different'. The internet, particularly with the advent of 3D printing, has created a global distribution platform that facilitates 'winner takes all' economies (Brynjolfsson & McAfee 2014): a small number of elite companies or individuals can capture entire markets with less need for human labour. This is coupled to accelerating advances in robotics and AI that make it more and more difficult for humans to reskill and retrain before the next bout of technological unemployment. Furthermore, even if humans and robots are equally good at work, there is a problem: robots are more efficient, less needy and untiring in their abilities. The logic of the profit motive will tend to favour robotic over human labour. I believe that there is good reason to think that the trend towards increasing technological unemployment is real. This does not mean that the future will see no human workers at all; just that it is likely to see a massive reduction in them.

But my purpose is not to fully convince you of this trend. Others have done that job more adequately and elaborately than I am able to do in this space. My purpose is to highlight the importance of the trend for the future of human moral agency. Although people do not always think of it in these terms, work is an important arena in which we can exert our moral agency. Work provides for the needs and wants of society, alleviating suffering and bestowing pleasure. It is how we earn money to provide for ourselves and our families; it is a space in which we develop moral virtues; and it is,

often, a source of great personal meaning and satisfaction. If this domain is taken away from us, one avenue for the development of our agency will be blocked. This is true even if we solve the distributional issues associated with the decline of paid employment. Many futurists and technophiles embrace the basic income as a solution to the problem of technological unemployment (e.g. Ford 2015). This would certainly alleviate some of the problems of technological unemployment by providing the disenfranchised worker with a source of income. But it is, in some ways, an agency-denying solution. If it is implemented, then we become more and more like moral patients upon whom the benefit of an income is bestowed; not moral agents who produce something of value from which we can earn an income.

In saying this, I do not wish to be taken to be endorsing some crude, capitalistic, pro-work worldview. I readily acknowledge that work can also be a source of misery and frustration for many, and that the need to work to secure a living may also deny and compromise our agency. Consequently, I accept that there are ways in which technological unemployment could actually enhance our moral agency by freeing us up to express it in more important arenas. This is the age-old response to those who worry about the displacing effects of technology. But this is only true if there are other more important arenas of agency that are immune from the rise of the robots. Two other technological trends give some reason for pessimism on this front.

The first of these two trends has to do with the rise of the robots within the political, legal and bureaucratic spheres. It would be wonderful if, having lost our jobs, we could take our new-found freedom and make it work to improve the political and legal infrastructure of our states. After all, this was, in many ways, the moral ideal of the Ancient Greeks. But with the rise of big data, the surveillance state, predictive analytics, and algorithm-based decision-making systems, the possibilities for human involvement in this sphere are also being hindered. These technologies allow for an increasing amount of automation in the implementation of legal rules and policies. We see this in how governments are already relying on data-mining algorithms to locate potential tax cheats, and facial recognition algorithms to locate cases of criminal fraud (Dormehl 2014). This trend is only set to grow. In fact, it is necessitated by the increasing technological complexity of the world that our laws try to regulate. For example, high frequency traders on Wall Street are not easily regulated by human beings. Machines are needed to assist in the task.

Now, you may accept that there is an increasing reliance on AI in the business of government, but dispute my claim that this will lead to a suppression of human agency. After all, it is unlikely that the desire for human elected officials and policy makers will disappear any time soon. The big data machines will simply be used to support the decisions made by these human agents. Indeed, you might even argue that agency could enhanced, not suppressed by their deployment: the machines will give human decision-makers better information upon which to make their moral decisions. This is an attractive claim, but it misses an important and subtle point. It is not that the increasing automation of legal and bureaucratic decision-making will completely eliminate the need for human beings to exercise decision-making authority; it is that it will reduce the scope for those humans to exercise their moral agency in making those decisions. If a machine-learning algorithm tells me that the right thing to do is to deny social welfare to someone on the grounds that they could be committing fraud, then it is true to say that I express some minimal form of agency by following through on that recommendation. But the agency in question really is minimal and not strongly moral. If I do not understand the rationale or basis on which the recommendation was made, and if I am not inclined or able to second-guess the algorithm's suggestion, I am little more than a 'rubber stamp'. I'm not exercising my capacity to understand the moral structure of the world or to make decisions on foot of that understanding. I'm leaving the algorithm do all the work. The claim I make here is that the use of big data systems and machine learning algorithms in public governance will result in exactly this type of 'rubber stamping' agency. The algorithms will detect patterns in the big datasets that humans would never be able to see; and they will issue policy recommendations that humans will never be able to challenge. At best, an elite of computer programmers and software designers will still have a meaningful role to play in the process; at worst, they too will see themselves surpassed by the machining learning algorithms they created. The result, once again, is a politico-legal infrastructure that suppresses agency and endorses patiency: we sit back and let the robots do all the work, occasionally clicking a button to implement their recommendations, and benefit from their ability to detect criminals and frauds and enforce the law.

Being shut out from the political and legal domain would be a significant loss of agency, but it would not leave us devoid of hope. With our new-found leisure time, we could simply enhance the expression of our agency in our personal lives. We could

pursue the humanistic and intellectual pleasures; enhance our personal fitness and well-being; seek out meaningful relationships with others; and produce works of artistic beauty. We would be more limited, to be sure, but we would still be able to express our agency. Indeed, our agency in this remaining domain could flourish.

I agree that there is some sliver of hope in this, but there is another trend in robotics and AI that threatens even this one remaining domain of agency. It is the rise of robotic and AI personal assistants, and the consequential outsourcing of our personal decision-making. This is already happening to our practices in dating and mating, wherein we frequently rely on algorithms to match us to other human beings; and in our personal health and fitness related activities, wherein in a variety of smart devices have been (and are being) created to motivate, cajole and reward us for our activities. There is still agency here, but there is an increasing outsourcing of the hard cognitive work needed to cultivate the agency-related virtues like courage and perseverance, and to understand the rational basis for our actions. And yes, to go back to Futurama for a moment, the growth of sex robots (and other pleasure bots) may simply exacerbate this suppression of agency.

4. Conclusion

To sum up, I have argued that the rise of robots and AI could pose a threat to the moral fabric of our civilization. It poses this threat because advances in these fields will have a tendency to suppress the expression of our capacity for moral agency and accentuate the expression of moral patiency. What's more, this threat does not stem from any fanciful speculation about all-powerful and superintelligent future forms of robots and AI. It stems from plausible extrapolations of existing trends. It will not lead to our extinction or annihilation — on the contrary, we will be the passive recipients of the benefits the technology can reap — but it will undermine a value that is foundational to our society.

There is, of course, one obvious objection to the argument I have just made (one that I have not yet considered). In talking about this threat and its effects, you may have noticed that I have taken on a somewhat fatalistic tone. I seem to be assuming that the trends in technological development are beyond our control: that, in some sense, the technology has a life of its own. But this, you could argue, is not true. We do have the

power to shape technology and we can use that power to prevent it from developing in ways that undermine our agency.

I would like to agree with this sentiment, and in some sense I hope that this argument serves as a self-undermining prophecy, but there is no sense in denying the difficulties. The development of robots and AI in manufacturing and service industries is driven by the seductive appeal of economic success; the growth of robots and AI in government stems from a commitment to generally-accepted values (efficiency, accuracy, cost-effectiveness) and from the need to respond to the complexity of the outside world; and, finally, our increasing willingness to outsource personal moral development to robots and AI is facilitated by psychological biases and heuristics. In other words, the trends stem from forces that are, to a considerable extent, larger and more powerful than us. Maybe there is room for some optimism. As I said earlier, technology won't rob us of our status as moral agents, merely suppress it. If we think agency is an important value, and we want to protect the value structure of contemporary civilization, we need to exercise it now.

Bibliography

Avent, R. (2016). *The Wealth of Humans*. London: St Martin's Press.

Bhuta, N., Beck, S., Geiβ, R., Liu, H-Y, and Kreβ, C (2016) *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: OUP

Brynjolfsson, E. and McAfee, A. (2014). *The Second Machine Age*. New York: WW Norton.

Calo, R., Froomkin, M., Kerr, I. (2016). *Robot Law*. Cheltenham: Edward Elgar Publishing.

Carr, N. (2014) *The Glass Cage*. London: The Bodley Head.

Danaher, J. (2014). Sex Work, Technological Unemployment and the Basic Income Guarantee. *Journal of Evolution and Technology* 24(1): 113-130.

Danaher, J. (2017). Robotic Rape and Robotic Child Sexual Abuse. *Criminal Law and Philosophy* 11(1): 71-95.

Dormehl, L. (2014). *The Formula: How Algorithms Solve all Our Problems…And Create More*. New York: Perigree.

Floridi, L. (1999). Information Ethics: On the Philosophical Foundation of Computer Ethics. *Ethics and Information Technology* 1(1): 37-56.

Ford, M. (2015). *The Rise of the Robots*. New York: Basic Books.

Frey, C and Osborne, M. (2013). The Future of Employment: How Susceptible are Jobs to Computerisation? *Oxford Martin School Working Paper*, September 2013 – available at http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

Griggs, B. (2014). Google's new self-driving car has no steering wheel or brake. *CNN* 24 May 2014, available at http://edition.cnn.com/2014/05/28/tech/innovation/google-self-driving-car/

Gunkel, D. (2011). *The Machine Question*. Cambridge, MA: MIT Press.

Hajdin, M. (1994). *The Boundaries of Moral Discourse*. Chicago: Loyola University Press.

Levy, D. (2007). *Love and Sex with Robots*. London: Harper Perennial.

Peter, F. (2008). Pure Epistemic Proceduralism. *Episteme* 5(1): 33-55

Susskind, R. and Susskind, D. (2015) *The Future of the Professions*. Oxford: OUP.

Van de Voort, M., Pieters, W., and Consoli, L. (2015). Refining the ethics of computer-made decisions: a classification of moral mediation by ubiquitous machines. *Ethics and Information Technology* 17(1): 41-56.