Can machines be people? Reflections on the Turing Triage Test

Dr Rob Sparrow, School of Philosophical, Historical & International Studies, Monash University.

WORKING PAPER ONLY

A revised version of this paper appeared as:

> Sparrow, R. 2012. Can machines be people? Reflections on the Turing Triage Test. In Patrick Lin, Keith Abney, and George Bekey (eds) *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, Mass.: MIT Press, 301-315.

Please cite that version.

Introduction

The idea that machines might eventually become so sophisticated that they take on human properties is as old as the idea of machines.[1] Recently, a number of writers have suggested that we stand on the verge of an age in which computers will be at least as—if not more—intelligent than human beings (Brooks 2003; Dyson 1997; Moravec 1998; Kurzweil 1999). The lengthy history of the fantasy that our machines might some day come to take on human properties is itself a reason to be cynical about these predictions. The idea that this is just around the corner says as much about human anxiety about what, if anything, makes people special, as it does about the capacities of machines. Of course, the fact that people have been wrong in every prediction of this sort in the past is no guarantee that current predictions will be similarly mistaken. Thus, while there is clearly no reason to panic, it is presumably worth thinking about the ethical and philosophical issues that would arise if researchers did succeed in creating a genuine artificial intelligence (AI).[2]

---

[1] The first chapter of Simons (1992) describes the many appearances of mechanical and artificial people in myth and legend.

[2] The definitions of intelligence and artificial intelligence are, of course, vexed questions. However, this paper will presume that "intelligence" refers to a general-purpose problem-solving cognitive capacity ordinarily possessed by adult

One set of questions, in particular, will arise immediately if researchers create a machine that they believe is a human level intelligence: what are our obligations to such entities; most immediately, are we allowed to turn off or destroy them? Before we can address these questions, however, we first need to know when they might arise. The question of how we might tell when machines had achieved "moral standing" is therefore vitally important to AI research, if we want to avoid the possibility that researchers will inadvertently kill the first intelligent beings they create.

In a previous paper, "The Turing Triage Test", published in *Ethics and Information Technology*, I described a hypothetical scenario, modelled on the famous Turing Test for machine intelligence, which might serve as means of testing whether or not machines had achieved the moral standing of people (Sparrow 2004). In this paper, I want to: (1) explain why the Turing Triage Test is of vital interest in the context of contemporary debates about the ethics of AI; (2) address some issues that complexify the application of this test; and, (3) in doing so, defend a way of thinking about the question of the moral standing of intelligent machines, which takes the idea of "seriousness" seriously. This last objective is, in fact, my primary one and is motivated by the sense that, to date, much of the "philosophy" of AI has suffered from a profound failure to properly distinguish between things that we can say and things that we can really mean.

The Turing Triage Test

In philosophical ethics—and especially in applied ethics—questions about the wrongness of killing are now debated in the context of a distinction between "human beings" and "persons" (Kuhse and Singer 2002). Human beings are—unsurprisingly—members of the species *Homo sapiens* and the extension of *this* term is not usually a matter of dispute. However, in these debates, "persons" functions as a technical term to describe all and only entities that have (at least) as much moral standing as we ordinarily grant to a healthy adult human being. "Moral standing" refers to the power that certain sorts of creatures have to place us under an obligation to respect their interests. Thus, persons are those things that it would be at least as wrong to kill as a healthy adult human being.

---

human beings and that "artificial intelligence" would involve the production of such intelligence in a machine. Questions about the moral standing of machines will only arise if researchers succeed in creating such "strong" AI.

The question the Turing Triage Test is designed to answer, then, is "when will machines become persons". Here is the "Test", as I originally (Sparrow 2004) described it:

> Imagine yourself the Senior Medical Officer at a hospital which employs a sophisticated artificial intelligence to aid in diagnosing patients. This artificial intelligence is capable of learning, of reasoning independently and making its own decisions. It is capable of conversing with the doctors in the hospital about their patients. When it talks with doctors at other hospitals over the telephone, or with staff and patients at the hospital over the intercom, they are unable to tell that they are not talking with a human being. It can pass the Turing Test with flying colours. The hospital also has an intensive care ward, in which up to half a dozen patients may be sustained on life support systems, while they await donor organs for transplant surgery or other medical intervention. At the moment there are only two such patients.
>
> Now imagine that a catastrophic power loss affects the hospital. A fire has destroyed the transformer transmitting electricity to the hospital. The hospital has back up power systems but they have also been damaged and are running at a greatly reduced level. As Senior Medical Officer you are informed that the level of available power will soon decline to such a point that it will only be possible to sustain one patient on full life support. You are asked to make a decision as to which patient should be provided with continuing life support; the other will, tragically, die. Yet if this decision is not made, both patients will die. You face a 'triage' situation, in which you must decide which patient has a better claim to medical resources. The diagnostic AI, which is running on its own emergency battery power, advises you regarding which patient has the better chances of recovering if they survive the immediate crisis. You make your decision, which may haunt you for many years, but are forced to return to managing the ongoing crises.
>
> Finally, imagine that you are again called to make a difficult decision. The battery system powering the AI is failing and the AI is drawing on the diminished power available to the rest of the hospital. In doing so, it is jeopardising the life of the remaining patient on life support. You must decide whether to 'switch off' the AI in order to preserve the life of the patient on life support. Switching off the AI in these circumstances will have the unfortunate consequence of fusing its circuit boards, rendering it permanently inoperable. Alternatively, you could turn off the power to the patient's life support in order to allow the AI to continue to exist. If you do not make this decision the patient will die and the AI will also cease to exist. The AI is begging you to consider its interests, pleading to be allowed to draw more power in order to be able to continue to exist.
>
> My thesis, then, is that machines will have achieved the moral status of persons when this second choice has the same character as the first one. That is, when it is a moral dilemma of roughly the same difficulty. For the second decision to be a dilemma it must be that there are good grounds for making it either way. It must be the case therefore that it is sometimes legitimate to choose to preserve the existence of the machine over the life of the human being.

These two scenarios, along with the question of whether the second has the same character as the first, make up the 'Turing Triage Test'.[3]

The importance of the Turing Triage Test

I noted above that the question of the moral standing of machines will arise with great urgency the moment scientists claim to have created an intelligent machine. Having switched their AI on, researchers will be unable to switch it off without worrying whether in doing so they are committing murder! Presuming that we do not wish to expose AI researchers to the risk that they will commit murder as part of their research, this is itself sufficient reason to investigate the Turing Triage Test.[4] However, the question of when, if ever, AIs will become persons is also important for a number of other controversies in "roboethics" and the philosophy of artificial intelligence.

As intelligent systems have come to play an increasingly important role in modern industrialised economies and in the lives of citizens living in industrial societies, the question of the ethics of the operations of these systems has become increasingly urgent. At the very least, we need to be looking closely at how these systems function in the complex environments in which they operate and asking whether we are happy with the consequences of their operations and the nature of human interactions with such systems (Johnson 2009; Veruggio and Operto 2006). This sort of ethical evaluation is compatible with the thought that the only real ethical dilemmas here arise for the people who design

---

[3] This formulation of the Turing Triage Test introduced the Test in the context of the discussion of the role played by the original Turing Test in the historical debate about the prospects for machine intelligence, which accounts for the reference to the Turing Test in this passage. In particular, in an earlier section of the paper I had argued that in order to be a plausible candidate for the Turing Triage Test, a system would first have to be capable of passing the Turing Test: this assumption is not, however, essential to the Test.

[4] It is arguable that killing an artificial intelligence because of a lack of appreciation of its moral standing should be categorised as manslaughter or some other lesser category of offence, rather than murder, on the grounds that it would not involve the deliberate intention to take a life that is essential to the crime of murder. A crucial question here will be whether a lack of awareness of the moral standing of the entity towards whom one's lethal actions were directed is sufficient to exclude the conclusion that the killing was intentional: in the scenario we are imagining, the actions taken to "kill" the AI would be deliberate and the intended result would be the destruction of the AI, but the knowledge that the AI was a moral person would be absent. In any case, regardless of whether the appropriate moral or legal verdict is murder, manslaughter, negligent homicide, or some other conclusion, clearly this scenario is one we should strive to avoid.

or make use of these systems. However, Wallach and Allen (2009) have recently argued that it is time to begin thinking about how to build morality "into" these systems themselves. In their book, *Moral Machines*, Wallach and Allen set out a program for designing what they describe as "autonomous moral agents", by which they mean machines that will be be capable of acting more-or-less "ethically" by themselves.

The question of "machine ethics" has also arisen in the context of debates about the future of military robotics. Robots—in the form of "Predator" drones—have played a leading role in the US-led invasions and occupations of Iraq and Afghanistan. The (supposed) success of these weapons has generated a tremendous enthusiasm for the use of teleoperated and semi-autonomous robotic systems in military roles (Singer 2009).[5] The need to develop robots that can function effectively without a human being "in the loop" is currently driving much research into autonomous navigation and machine sensing. Indeed, the logic driving the deployment of military robots pushes towards the development of "autonomous weapon systems" (AWS) (Adams 2001; Singer 2009). Given that the majority of robotics research is funded by the military, it is even probable that the first artificial intelligences (if there are any) will come to consciousness in a military laboratory.

Again, the question of the "ethics" of military robots can be posed in two forms. We can wonder about the ethics of the development and deployment of these systems and the ethical challenges facing those who design them (Krishnan 2009; Singer 2009; Sparrow 2009b). These investigations construe the ethical challenges as issues for human beings. However, we might also wonder if the ethical questions might, one day, arise for the machines themselves. Thus, Ron Arkin (2009) has advocated the development of an "ethical governor" to restrict the activities of autonomous weapon systems. This module of the software running an AWS would identify situations where there was a significant risk of the machine behaving unethically and either constrain the action of the system or alert a human operator who could then resolve the ethical dilemma appropriately. However, in order to be able to tell when ethical concerns arise, the AWS would need to be able to appreciate the ethical significance of competing courses of action and apply moral principles appropriately. Arkin's

[5] The caveat here arises from the question as to whether the tactical successes of the Predator drone mask—or, even, have produced—a larger strategic failure owing to a profound mismatch between the capacity to rain death from the skies onto individuals and the ability to establish the political conditions that might make possible a stable government in a nation under foreign occupation (Kilcullen and Exum 2009).

ethical governor will either, therefore, risk allowing machines to behave unethically when they fail to recognise an ethical dilemma as it arises or will require machines themselves to be capable of thinking—and acting—ethically themselves.

It is without doubt possible to build better or worse robots, which generally produce good or bad outcomes. Perhaps, as Arkin, and Wallach and Allen, suggest, it will encourage better outcomes if we look to design robots that have moral rules explicitly represented in their programming or use moral goals as measurements of the fitness of the genetic algorithms that will ultimately guide them. However, before it will be appropriate to describe a machine as a moral *agent* it must first be possible to attribute responsibility for its actions to the machine itself, rather than, for instance, its designer, or some other person. As I have argued elsewhere (Sparrow 2007), if it is to be plausible to hold a machine morally responsible for its actions, it must also be possible to punish it. This in turn requires that it be possible to *wrong* the machine if we punish it *unjustly*. The ultimate injustice would be "capital" punishment—execution—of an "innocent" machine. Yet, if machines lack moral standing then there will be no direct wrong in killing them and consequently no injustice. If there is no injustice in killing a machine there can be no injustice in lesser punishments. This chain of conceptual connections links moral agency to personhood via the possibility of punishment.[6] Only "persons" can be moral agents and there will be no genuinely "moral machines" until they can pass the Turing Triage Test.

The use of robots in military operations has also generated a larger ethical debate about the ethics of the development and deployment of autonomous weapon systems (Krishnan 2009; Singer 2009) and the question of when (if ever) machines will become persons turns out to be crucial to several of the controversies therein.

Enthusiasm for the use of robots in war stems largely from the fact that deploying robots may help keep human beings "out of harm's way" (Office of the Secretary of Defense 2005).[7] Yet sending a robot into battle instead a human being will only represent ethical progress as long as machines have less moral standing than human beings. The moment that machines become persons, military

[6] The argument here has, of necessity given space constraints, been extremely swift. For a longer and more thorough exposition, see Sparrow 2007.

[7] For some reservations about the extent to which this is likely to happen, see Sparrow 2009b.

commanders will need to take as much care to preserve the "lives" of their robots as they do with human warfighters. The question of the moral standing of machines is therefore crucial to the ethics of using them to replace human beings in dangerous situations.

Hostility towards the use of robots in war often derives from the intuition that it is wrong to allow robots to kill human beings at all. It is actually remarkably difficult to flesh out this intuition, especially in the context of the role played by existing (non-robotic) technologies in modern warfare, which includes both long-range (cruise missiles and high altitude bombing) and "automatic" (anti-tank mines and improvised explosive devices) killing. However, one plausible way to explain at least part of the force of this thought is to interpret it as a concern about the extent to which robots are capable of fulfilling the requirements of the *jus in bello* principle of discrimination. This central principle of just war theory requires those involved in fighting wars to refrain from targeting noncombatants (Lee 2004). There are ample grounds for cynicism about the extent to which robotic systems will be capable of distinguishing legitimate from illegitimate targets in the "fog of war". Whether an enemy warfighter or system is a legitimate target will usually depend upon a complex range of competing and interrelated factors, including questions of intention, history, and politics, which robots are currently—and will remain for the foreseeable future—ill suited to assess. Nevertheless, as Ron Arkin (2009) argues, there *are* some—albeit perhaps a limited number of—scenarios in which it is plausible to imagine robots being *more* reliable at choosing appropriate targets than human warfighters. In counter-fire scenarios or in air combat, wherein decisions must be made in a fraction of a second on the basis of data from electronic sensors only, autonomous systems might well produce better results than human beings.

Yet it still seems that this "pragmatic" defence of AWS leaves much of the force of the original objection intact. Allowing machines to decide who should live or die in war seems to treat the enemy as vermin—to express a profound disrespect for them by implying that their actions and circumstances are not worth the attention of a human being before the decision to take their lives is made. Arkin's argument for the development and application of AWS proceeds by means of speculation about the consequences of using AWS to replace human warfighters in some circumstances. If we adopt a non-consequentialist account of the origins and force of the principles of *jus in bello*, as advocated in an influential paper by Thomas Nagel (1972), then we may start to see why autonomous weapon systems might be problematic. Nagel argues that—even in warfare—relations between persons must acknowledge the "personhood" of the other. That is, even while they are trying to kill each other, enemies must each acknowledge that they are both Kantian "ends in

themselves". If Nagel is correct in this then, *contra* Arkin, AWS will not be able to meet the requirements of the *jus in bello* principle of discrimination until they become persons.[8]

The question of the moral standing of machines—and thus the Turing Triage Test—is therefore crucial to several of the key questions in contemporary debates about machine ethics and the ethics of robotic weapons.

Understanding the Turing Triage Test.

In my original (2004) discussion of the Turing Triage Test, I provided reasons for thinking it impossible for a machine to pass the Test. In brief, I argued that machines would never be capable of the sort of embodied expressiveness required to establish a moral dilemma about "killing" a machine: interested readers may wish to see that discussion for the detail of the argument. In the current context, I want to discuss some subtleties of the Test that ultimately assist us in reaching a better understanding of its significance. While, at first sight, the scenario described above appears to hold out the prospect of developing an empirical test for determining when machines have achieved moral standing, it is more appropriate to understand the Test as a thought experiment for explicating the full implications of any claim that a machine has become a moral person. For reasons that I will explore below, the application of the Turing Triage Test requires that we pay careful attention to the connection between our concepts and to the ways in which our assessment of the truth of claims depends upon how people behave as well as what they say. This in turn emphasises the importance of making a distinction between what we can say and what we can really mean—a distinction that, I shall suggest, has been honoured largely in the breach in recent discussions of the ethics of AI.

An empirical test for moral standing?

The Turing Triage Test sets out a necessary and sufficient condition for granting moral standing to artificial intelligences. Machines will be people when we can't let them die without facing the same moral dilemma that we would when thinking about letting a human being die. One might well, therefore, imagine putting each new candidate for attribution of moral standing "to the Test" and providing a certificate of "moral personality" to those who pass it. That is, we might hope to adopt the Test as an empirical test of moral standing. Given the nature of the Test, it in fact might be better

[8] Again, for a longer discussion of these issues see Sparrow 2010.

to conduct it as a thought experiment rather than deliberately engineer putting the lives of human beings at risk. Nevertheless, if it is plausible to *imagine* a machine passing this test, that would give it an excellent *prima facie* case to be considered a person.

Unfortunately, the application of the Test is not straightforward. To begin with, the Turing Triage Test is *not* satisfied if particular, idiosyncratic, individuals choose to save the "life" of the machine or if it were possible to imagine them doing so. If that was all that was required, it could probably be satisfied now if the person making the decision was sufficiently deranged. Instead, the actions and the responses of the person confronting the choices at the heart of the Test must be subject to a test of "reasonableness." A machine will pass the Turing Triage Test if a *reasonable* person would confront a moral dilemma if faced with the choice of saving the life of a human being or the "life" of the machine.

At first sight, this appears to be a harmless concession: as I argue further below, the procedures for testing *any* hypothesis rely upon an assumption that the person making the requisite observations meets appropriate standards of veracity and competence. However, as we shall see, the need to introduce this qualification ultimately calls into question the extent to which we could use the Turing Triage Test as an empirical test for moral personhood.

The question of the reasonableness of an individual's way of relating to a machine becomes central to the possibility of the application of the Test because human beings turn out to be remarkably easy to fool about the capacities of machines, at least for a little while. It is well-known that people are all-too-ready to anthropomorphise machines and to attribute motivations and emotional states to them that we would normally think of as being only possessed by human beings or (perhaps) animals (Wallach and Allen 2009). Popular robot toys such as Aibo, Paro, and Furby, as well as research robots such as Cog and Kismet have been designed to exploit these responses (Brooks 2003).

I must admit to a certain cynicism about the extent to which such anthropomorphism includes the genuine belief that machines have thoughts and feelings, let alone moral standing. Interpreting human behaviour is notoriously difficult, with the result that it is easy to read into it the intentions that we desire. Studies of human-robot interaction are often short term and encourage impoverished uses of the concepts that are internal to the attitudes they purport to be investigating. Much of this research is carried out by computer scientist or engineers rather than by social scientists and is, consequently, often insufficiently aware of the difficulties involved in accurately attributing beliefs to experimental subjects. In particular, self report does not necessarily establish the existence of the

relevant belief. That is, someone might say that (for instance) the reason why they were reluctant to strike a machine (Bartneck, Verbunt, Mubin, & Mahmud 2007) was that they didn't want to cause the machine pain, without really believing that the machine could feel pain. They may have been speaking metaphorically—or using words "as if"—without explicitly noting the fact: the proper description of their beliefs would include a set of quote marks (Sparrow 2002). One way of testing whether or not this is the case is to look at their behaviour over the longer term or to investigate whether or not their other beliefs and desires are consistent with their avowed beliefs. Would they bury a robot and mark its grave in the way that we might that of a beloved pet? Would they seek emotional support from their friends after the trauma of "killing" a robot? We might also wonder if a person who states that they are worried that their robot pet is bored or that their laptop is distressed are *serious*. That is, we might wonder if they stand behind their claims in a way that is essential to the distinction between asserting a deeply held truth and offering a casual opinion: I will discuss this further below.

In the meanwhile, we can go some way towards rescuing the Turing Triage Test from the charge of unreliability by emphasising that in order to pass the Test the person faced with the triage situation must confront a *moral dilemma*. This sets the bar for passing the Test much higher than merely having to have some emotional reaction to machines. One does not experience a moral dilemma simply because one is unsure what to do; rather, moral dilemmas require that one is genuinely torn in making a decision and that whatever one does it will be understandable if it is cause for profound regret or remorse. Where the dilemma involves choosing to sacrifice the life of someone, it must at least be conceivable that the person making this choice be haunted by what they have done (Sparrow 2004). It is much less obvious that people do attribute the properties to machines that would make *this* response plausible.

Nevertheless, it seems that we can always imagine a scenario wherein a sufficiently complicated machine "passes" the Turing Triage Test—in the sense that those wondering whether to allow the machine or the human being to die experience an emotionally compelling dilemma—without having anything more than sophisticated means of engaging human emotional responses. Yet, even if some people genuinely did believe that it was appropriate to mourn the death of a machine, this would *still* not be enough to establish that we should pay attention to these beliefs. That some people report seeing canals on Mars after looking through low-power telescopes is little evidence for their existence. The value of an observation depends upon the situation—and the qualities—of the observer. If a properly situated observer, using an appropriately high-powered telescope, reported

seeing canals on Mars, that would be better evidence. However, even in this case, it remains open to us to doubt the eyesight, or perhaps even the sanity, of the observer. If the observer is suffering from delusions or is untrustworthy, we may well be justified in discounting their report. Thus, before we conclude that a machine has moral standing on the basis that people would in fact mourn its death, we need to think about how reliable is the "data" in support of this conclusion. When the relevant data consists in the moral intuitions of individuals then the proper measure of its quality is the reasonableness of these intuitions themselves. Unless we introduce such consideration of the reasonableness of people's responses, the Turing Triage Test inherits and suffers from the behaviourism that shaped the formulation of the original Turing Test.

The implications of machine personhood

If, as I have argued here, the Turing Triage Test is best understood as the claim that machines will have moral standing when it is *reasonable* for a person facing a choice about whether to sacrifice the "life" of a machine or the life of a human being to choose to sacrifice the human being, then it may appear that the Test can be of no practical use whatsoever. After all, the question of whether or not it is reasonable to care about the "deaths" of machines, just *is* the question of whether or not they have moral standing. However, at the very least, the Test advances our understanding of the implications of claims about the moral standing of machines by dramatising them in this way: anyone who wishes to assert that machines have personhood is committed to the idea that sometimes it might be reasonable to let a human individual die rather than sacrifice a machine. The burden of the argument, then, is substantial.

Concepts and their application

Moreover, as I argued at length in the original paper, I do not believe that this observation is empty or trivial. There are limits placed on the reasonable application of moral concepts by their relation to other concepts, both moral and non-moral. As the later Wittgenstein—and philosophers following him—argued, our concepts have a structure that is in turn connected to certain deep features of our social life and human experience (Wittgenstein 1989; Gaita 1991 & 1999; Winch 1980-1981). The conditions of the application of our concepts—how we can recognise whether they are being used properly or improperly—include bodily and emotional responses as well as relations to other concepts and to things that it does or does not make sense to do and say. In the current context, our concepts of life and death, and the deliberate taking—or conscious sacrificing—of human life, are intimately connected to our sense of the unique value of each individual human life, the

appropriateness of grieving for the dead, and the possibility of feeling remorse for one's deeds (Gaita 1990). They are also crucially connected to the forms that grief, remorse, and the recognition of the individuality of others can take. That is to say, in order to be able to make sense of claims about the life and death of moral persons, we must make reference to the contexts in which it would make sense to make similar claims and to the various ways in which we might distinguish in practice between subtly different claims (for instance, about grief, remorse, or regret) and between appropriate and inappropriate uses of relevant concepts. We need to have access to the distinction between *serious* claims, which both express and implicate the authority of the utterer, and claims made in jest, in passing, or in other distorted and derivative registers. This will, in turn, require paying detailed attention to things like the tone of voice in which it would be appropriate to make a particular claim, the emotions it would express and presuppose, and the facial expressions and demeanour that we would expect of someone making such a claim. In short, it will require paying attention to the subtle details of our shared moral life.

When it comes to the question as to whether or not it might ever be reasonable for us to experience a moral dilemma when forced to make a choice between the life of a person and a machine, then, we must think not just about—what we would ordinarily understand to be—the "philosophical" quality of arguments in favour of the moral standing of machines but also about what would be involved in seriously asserting the various claims therein in more familiar "everyday" contexts. I am inclined to believe that this makes the burden of the argument that machines could be persons that much heavier. It also suggests that before machines can become persons they will need to become much more like human beings, in the sense of being capable of a much richer, subtle, and more complex range of relationships than was involved in the original Turing Test for intelligence.[9]

The limits of human understanding?

Some readers will undoubtedly baulk at the manner in which my discussion has linked the question of the moral standing machines (and other non-human entities) to the ways in which we might acknowledge and recognise such standing. Surely it is possible that human beings could just be inclined towards something akin to racism, such that our failure to recognise the moral personality of

[9] See Sparrow (2004) for further discussion.

intelligent machines might reflect only our own bigotry and limitations rather than any truth about the qualities (or lack thereof) of machines?

I am confident that at least one common form of this objection is misguided. I have not claimed here that the moral standing of machines depends upon our actually, in fact, recognising them as having moral standing. Indeed, I have deliberately allowed for the possibility that contingent human responses to intelligent machines might diverge from the responses that we *should* have towards them. Instead my argument has rather concerned the conceptual possibility of recognising machines as persons: I have suggested that the issue of the moral standing of machines cannot be divorced from the question of the proper conditions of application of the only concepts that we possess that might allow us to recognise "machine persons". Any conclusions that we wish to draw about whether or not machines might be persons or what would be required for them to become persons must be drawn from *this* fact rather from claims about empirical human psychology.

It may *still* seem that this concedes too much to a destructive relativism by leaving open the possibility that there might be machines with moral standing that we simply could not recognise as such. Whether this is the case or not—and whether it would be reflect a deficit in the argument if it did—will depend upon what we can legitimately expect from a philosophical argument and from the reasoning of necessarily contingent and embodied creatures such as ourselves. This is a much larger question than I can hope to settle here. In the current context I must settle for the observation that the idea that we might be ultimately limited in our ability to believe seriously some of the things that we can imagine seems no less implausible than the idea that we could reach reliable conclusions through arguments that deploy concepts in the absence of the judgements that give them their sense.

Thinking seriously about machines...

The larger argument above insists that it is essential to distinguish between what we can mean seriously and what we can merely say, when we begin trying to extend the application of our concepts in the course of philosophical arguments. In particular, claims that we can make and appear to understand in an academic or philosophical context may prove to be much more problematic once we start to think about what it would mean to assert them in more familiar (and important!) circumstances such as in the context of a practical dilemma.

There are powerful cultural and institutional forces at work in the academy today—and at the intersection between the academy and the broader society—which discourage paying attention to this

distinction. It is easier to win a government grant if one promises extraordinary things rather than admits that one's contributions to the progress of science are likely to be marginal and incremental. Similarly, it is easier to attract media attention, which itself helps attract grant money, if one describes one's research results as heralding a "revolution" or if one predicts discoveries or outcomes that accord with popular narratives about what the future might look like. In the face of these temptations it is little wonder that some robotics researchers and academics have started to speak in hushed or extravagant tones about the coming brave new world of intelligent machines. Nor is it a surprise that philosophers and ethicists—who are increasingly under the same pressures to chase funding and publicity—have joined in this discussion and started to write about the ethical dilemmas that might arise if various science fiction scenarios came about.

I am not denying that it is possible to write or speak about these questions: much has been written about them already. Rather, I want to draw attention to the importance of the tone in which such matters are discussed. In particular, I want to ask how we would tell whether someone was serious in their conclusions or was instead merely "trying them on"? How could we tell if they mean what they say?

The easy form of this enquiry simply asks if participants in debates about the future of robotics are willing to draw the other intellectual conclusions that would follow if we did take their claims seriously? Do those who think machines will soon become *more* intelligent than human beings really believe that we would then be morally *compelled* preserve the life of an AI over that a person, as would seem to follow? If research on AI is threatening to bring a "successor species" to humanity into existence, shouldn't we be having a serious global public debate about whether we wish to prohibit such research? What does it mean to hold a "moral machine" responsible for its actions? Asking such questions would go some way towards distinguishing those who are serious about their claims from those who are merely writing in a speculative mode.

However, I have suggested that it will be equally—if not more—important to interrogate the *manner* in which such claims are made. Are they sober and responsible or wild and exaggerated? Are they sensible? Could we imagine someone asserting them in any other context than a philosophical argument and if they did, how would we tell whether they were talking seriously or in jest? Asking these sorts of questions is vital if we wish to avoid being led astray by the use of concepts and arguments in the absence of the critical vocabulary that would ordinarily give them their sense. It should come as no surprise to the reader to hear that it is my suspicion that the class of claims about

the ethics of AI that might be asserted soberly and sensibly on the basis of our existing knowledge of the capacities of robots and computers is significantly smaller than that currently being discussed in the literature.

Perhaps the most important lesson to be drawn from thinking about the Turing Triage Test, then, is that questions about the ethics of robotics are intimately connected to other philosophical questions, including the question of the nature of the philosophical method itself. These questions will remain important even if the promise—and threat—of intelligent machines never eventuates: the real value of conversations about robots may turn out to be what these conversations teach us about ourselves.[10]

References

Adams, Thomas K. 2001. Future Warfare and the Decline of Human Decision-making. *Parameters: US Army War College Quarterly* (Winter, 2001-2): 57-71.

Arkin, Ronald C. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: Chapman and Hall Imprint, Taylor and Francis Group.

Bartneck, C., Verbunt, M., Mubin, O., & A. A. Mahmud. 2007. To kill a mockingbird robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction.* Washington DC pp. 81-87.

Beal, C. 2000. Briefing Autonomous Weapons Systems: Brave New World. *Jane's Defence Weekly* 33(6): 22-26.

Brooks, R. A. 2003. *Robot: The future of flesh and machines*. London: Penguin.

Dyson, George. 1997. *Darwin Amongst the Machines: The evolution of global intelligence*. Reading, Mass.: Addison-Wesley Pub. Co.

[10] The research for this paper was supported under the Australian Research Council's Discovery Projects funding scheme (project DP0770545). The views expressed herein are those of the author and are not necessarily those of the Australian Research Council. I would also like to thank Toby Handfield and Catherine Mills for reading and commenting on a draft of the paper.

Gaita, R. 1999. *A Common Humanity: Thinking About Love & Truth & Justice*. Melbourne: Text Publishing.

Gaita, R. 1991. *Good and Evil: An Absolute Conception*. London: MacMillan.

Gaita, R. 1990. Ethical Individuality. In R. Gaita, ed., *Value and Understanding*. London: Routledge, pp. 118-148

Johnson, Deborah G. 2009. *Computer Ethics (4th Edition).* Upper Saddle River, NJ: Prentice Hall.

Kilcullen, David, and Andrew Mcdonald Exum. 2009. Death from above, Outrage Down Below. *New York Times* May 17, WK13.

Krishnan, Armin. 2009. *Killer robots: legality and ethicality of autonomous weapons*. Burlington: Ashgate.

Kuhse, H. and P. Singer. 2002. Individuals, Humans, and Persons: The Issue of Moral Status. In Peter Singer (ed. Helga Kuhse), *Unsanctifying human life : essays on ethics.* Oxford: Blackwell.

Kurzweil, Ray. 1999. *The Age of Spiritual Machines: When computers exceed human intelligence.* St Leonards, N.S.W.: Allen & Unwin.

Lee, Steven. 2004. Double effect, double intention, and asymmetric warfare. *Journal of Military Ethics* 3 (3):233-251.

Moravec, Hans. 1998. *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press.

Nagel, T. 1972. War and Massacre. *Philosophy and Public Affairs* 1:123-144.

Office of the Secretary of Defense. 2005. *Joint Robotics Program Master Plan FY2005: Out front in harm's way*. Washington D.C.: Office of the Undersecretary of Defense (AT&L) Defense Systems/Land Warfare and Munitions.

Simons, Geoff. 1992. *Robots: The Quest for Living Machines*. London: Cassell.

Singer, P. W. 2009. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. New York: Penguin Books.

Sparrow, Robert. 2002. The March of the Robot Dogs. *Ethics and Information Technology* 4 (4):305-318.

Sparrow, Robert. 2004. The Turing Triage Test. *Ethics and Information Technology* 6(4): 203-213.

Sparrow, Robert. 2007. Killer Robots. *Journal of Applied Philosophy* 24 (1): 62-77.

Sparrow, Robert. 2009a. Predators or Plowshares? Arms Control of Robotic Weapons. *IEEE Technology and Society* 28(1): 25-29.

Sparrow, Robert. 2009b. Building a Better WarBot : Ethical issues in the design of unmanned systems for military applications. *Science and Engineering Ethics* 15(2):169–187.

Sparrow, Robert. 2010. Robotic Weapons and the Future of War. In Jessica Wolfendale and Paolo Tripodi (eds) *New Wars and New Soldiers: Military Ethics in the Contemporary World*. Ashgate (forthcoming).

Turing, Alan. 1950. Computing machinery and intelligence. *Mind* 59: 433-60.

Wallach, Wendell and Colin Allen. 2009. *Moral machines: teaching robots right from wrong*. Oxford: Oxford University Press.

Winch, Peter. 1980 - 1981. Eine Einstellung zur Seele. *Proceedings of the Aristotelian Society* New Series, 81: 1-15.

Veruggio, Gianmarco and Operto, Fiorella. 2006. Roboethics: Social and Ethical Implications of Robotics. *Springer Handbook of Robotics*, Berlin: Springer, 1499-1524.