The Turing Triage Test

Dr. Robert Sparrow

Centre for Applied Philosophy and Public Ethics

Working Paper Number 2002/14

Centre for Applied Philosophy and Public Ethics (CAPPE)

CAPPE Melbourne
Department of Philosophy
University of Melbourne
Parkville, Victoria, 3010
Phone: (03) 9344-5125
Fax: (03) 9348-2130

CAPPE Canberra
GPO Box A260
Australian National University
Canberra, 2601
Phone: (02) 6125-8467
Fax: (02) 6125-6579

The Centre for Applied Philosophy and Public Ethics Working Paper Series

The Centre for Applied Philosophy and Public Ethics (CAPPE) was established in 2000 as a Special Research Centre in applied philosophy funded by the Australian Research Council. It has combined the complementary strengths of two existing centres specialising in applied philosophy, namely the Centre for Philosophy and Public Issues (CPPI) at the University of Melbourne and the Centre for Professional and Applied Ethics at Charles Sturt University. It operates as a unified centre with two divisions: in Melbourne at the University of Melbourne and in Canberra at Charles Sturt University. The Director of CAPPE and the head of the Canberra node is Professor Seumas Miller. Professor C.A.J. (Tony) Coady is the Deputy Director of CAPPE and the head of the Melbourne node.

The Centre concentrates, in a single unit, the expertise of applied philosophers working in diverse fields of research. The Centre promotes community discussion and professional dialogue concerning key ethical problems facing Australia today. It is Australia's leading centre of research excellence in this area and it has extensive links with international institutions and scholars. The Centre has also established collaborative projects with a number of Australian and overseas universities. The Melbourne division of the Centre, in its previous form as CPPI, has conducted business ethics consultancies and roundtables with many of Australia's leading companies, including Shell, Telstra, BHP, Sydney Water and Western Mining. Such activities continue in CAPPE - it sponsors workshops, conferences and public lectures on topics of current public interest.

The Occasional Paper Series was developed by CAPPE as a forum for the dissemination of ideas prior to publication in academic journals or books. It is hoped that publication in this series will elicit constructive criticisms and comments from specialists in the relevant field. Inclusion in this series does not preclude publication elsewhere.

Michael O'Keefe, Series Editor

Draft Only: Not to be cited without permission.

The Turing Triage Test

Robert Sparrow

Abstract

If, as a number of writers have predicted, the computers of the future will possess intelligence and capacities that exceed our own then it seems as though they will be worthy of a moral respect at least equal to, and perhaps greater than, human beings. In this paper I propose a test to determine when we have reached that point. Following Alan Turing's (1950) original 'Turing test', which argued that we would be justified in conceding that machines could think if they could fill the role of a person in a conversation, I propose a test for when computers have achieved moral standing by asking when a computer might take the place of a human being in a moral dilemma, such as a 'triage' situation in which a choice must be made as to which of two human lives to save. We will know when machines have achieved moral standing comparable to a human when the replacement of one of these people with an artificial intelligence leaves the character of the dilemma intact. That is, when we might sometimes judge that it is reasonable to preserve the continuing existence of a machine over the life of a human being. This is the 'Turing Triage Test'. I argue that if personhood is understood as a matter of possessing a set of important cognitive capacities then it seems likely that future AIs will be able to pass this test. However this conclusion serves as a reductio of this account of the nature of persons. I set out

an alternative account of the nature of persons, which places the concept of a person at the centre of an interdependent network of moral and affective responses, such as remorse, grief and sympathy. I argue that according to this second, superior, account of the nature of persons, machines will be unable to pass the Turing Triage Test until they possess bodies and faces with expressive capacities akin to those of the human form.

Keywords: Artificial Intelligence, computers, ethics, Turing Test, person, embodiment.

The Turing Triage Test

INTRODUCTION

If we are to believe the pronouncements of some researchers in the field of artificial intelligence, it will not be long until computers become autonomous systems, making decisions on their own behalf. In the not too distant future, computers will have beliefs and desires, even emotions, in order that they can reason better and function in a wider range of situations. They may even 'evolve' via genetic algorithms, genetic programming or other methods of evolutionary computation. Eventually, through these techniques or simply through increasingly sophisticated design, they will become fully fledged self-conscious 'artificial intelligences'. According to a number of writers in the field, before the end of the century – and according to some, well before this – machines will be conscious, intelligent, entities with capacities exceeding our own (Dyson, 1997; Moravec, 1988; Moravec, 1998; Kurzweil, 1992; Kurzweil, 1999; Simons, 1992).

As soon as AIs begin to possess consciousness, desires and projects then it seems as though they deserve some sort of moral standing.[1] For instance, if my computer has more intelligence than my dog, is self conscious and has internal states that function as pleasure and pain, and hopes and dreams, then it seems as though it would be at

least as wrong to destroy it as to kill my dog. If, as a number of writers have predicted, artificial intelligences will eventually possess intelligence and capacities that exceed our own then it seems as though they will be worthy of a moral respect at least equal to, and perhaps greater than human beings. We may have duties towards such entities in our relations with them. It may even become necessary to grant them rights comparable to those possessed by human beings.

In this paper I propose a test to determine when we have reached that point. Following Alan Turing's (1950) original 'Turing test', which argued that we would be justified in conceding that machines could think if they could fill the role of a person in a conversation, I propose a test for when computers have achieved moral standing by asking when a computer might fill the role of a human being in a moral dilemma. The dilemma I have chosen is a case of 'triage', in which a choice must be made as to which of two lives to save. In the scenario I propose, a hospital administrator is faced with the decision as to which of two patients on life support systems to continue to provide electricity to, following a catastrophic loss of power in the hospital. She can only preserve the existence of one and there are no other lives riding on her decision. We will know when machines have achieved moral standing comparable to a human when the replacement of one of the patients with an artificial intelligence leaves the character of the dilemma intact. That is, when we might sometimes judge that it is reasonable to preserve the continuing existence of the machine over the life of the human being. This is the '**Turing Triage Test**'.

SOME QUALIFICATIONS

‘Weak’ versus ‘Strong’ AI

Before I proceed with my discussion, I wish to head off an objection that might be made by those who would argue that I am misrepresenting the nature of research into artificial intelligence. Some researchers into advanced computing have given up the attempt to create artificial intelligences of the sort that I will be discussing. They have concluded either that the creation of genuine intelligence is beyond our current technological prowess or that there exists no single human capacity of intelligence that might be artificially reproduced. Instead they dedicate themselves to designing machines that can perform tasks similar to those performed by the human brain in some more narrowly prescribed area, such as facial or speech recognition, vision, or problem solving of certain sorts. Projects of this type are often described as ‘Weak AI’. Typically, researchers involved in Weak AI wish to avoid the question as to whether success in these endeavours might ever involve the creation of genuine intelligence. To talk of machines, having ‘intelligence’, let alone ‘beliefs and desires’ or ‘self consciousness’, is to confuse appearance with reality and, what’s more, to risk provoking a dangerous backlash against their research by fuelling the public’s perception that they are modern day Frankensteins. What are misleadingly described in the popular press as ‘artificial intelligences’ are simply more complicated machines

that are capable of performing complex tasks that in the past have only been possible for human beings.

As will become clear below, I have some sympathy for this position's dismissal of the possibility of genuine artificial intelligence. It may be that the technology never achieves the results necessary to create the issues with which I am concerned here.[2] But despite the lowered sights of some 'AI' researchers, other researchers do claim to be working towards the creation of genuine artificial intelligence – a project known as 'Strong AI'. This paper takes the optimistic rhetoric of Strong AI enthusiasts at face value, at least initially; after all, what if they are right? It is best if we start talking about the ethical dilemmas now. Furthermore, it is dangerously presumptuous to claim that science will never progress to the point at which the question as to the moral status of intelligent computers arises. Computer engineers and scientists have in the past shown a marked ability to disconcert the pundits by greatly exceeding expectations and achieving results previously thought impossible. If they do succeed in creating genuine artificial intelligence then the issue of the range and nature of our obligations towards them will arise immediately.

Artificial Intelligence and the 'Turing Test'

Before I continue then, I need to say something about what I mean by 'artificial intelligence'. The definition of intelligence is a vexed question in the philosophy of mind. We seem to have a firm intuitive grasp of what intelligence is. Roughly

speaking, it is the ability to reason, to think logically, to use imagination, to learn and to exercise judgement. It is the ability to frame a problem and then solve it. Intelligence is generalisable; it is capable of doing these things across a wide range of problems and contexts. It is what we have, what primates have less of, parrots still less, jelly fish and trees (and contemporary machines) not at all. Artificial intelligence is intelligence in an artefact that we have created.

Yet it is surprisingly difficult to give a complete description of what intelligence consists in, let alone a precise definition. Because of the difficulty of providing a definition of intelligence, much of the discussion in the AI literature has moved to the hopefully easier question of how we might **tell** whether a machine was intelligent, even if we are unsure of exactly what intelligence consists in. This discussion has largely focussed on the appropriateness or otherwise of the notorious 'Turing Test'. In his famous article in Mind, Alan Turing (1950) suggested that we would be justified in conceding that machines were intelligent if they could successfully take the part of a human being in a conversation over a teletype machine. If we cannot tell the difference between a human being and a machine in the course of a conversation in the absence of visual cues, then we must acknowledge that the machine is intelligent.[3]

The adequacy of this test for machine intelligence has been the subject of controversy ever since. Critics have alleged that the Turing Test sets the standard for intelligence too high, too low, or in the wrong place altogether.[4] I cannot enter these

debates here. I can only state my belief that, if anything the Turing Test sets the standard of behaviour for intelligence too high (after all, chimpanzees are intelligent - to a degree at least - and cannot pass the test). In any case, for the purposes of this paper I shall assume that whether or not an ability to pass the Turing Test is a necessary condition for possession of intelligence it is at least a sufficient condition.[5]

Furthermore, I will also assume that the Turing Test also establishes more than is usually claimed on its behalf. I will hold that a machine that can pass the Turing Test should be acknowledged to be self-conscious as well as intelligent and also to have projects and ambitions that matter to it. While Turing himself did not argue that his test would establish these further conclusions, there is at least a prima facie case that it should. If a machine is to be able to converse like a human being then it must be capable of reporting on its internal states and its past history. That is, it must be able to demonstrate an awareness of self. Questions about our feelings and our personal history are a natural part of conversation. It is difficult to imagine how a machine which did not possess self-consciousness could carry on a convincing conversation about these things. Similarly, the fact that we have hopes and dreams, projects and ambitions, and that these matter to us, is also something that is evidenced in conversation. For instance we ask often each other about our intentions, ambitions, and attitudes towards various events and circumstances in our lives. We express happiness and sadness, joy and anger, concerning the satisfaction or frustration of our desires in the course of conversation. A machine which was unable to do the same would be unable to pass the Turing Test.

A machine that can pass the Turing Test must therefore be able to behave as though it has self-consciousness and commitments to various projects. If successful imitation of intelligent behaviour is sufficient to establish the presence of intelligence than so too, I argue, should it establish the presence of these further capacities.[6] Of course this argument is controversial; as is, for that matter, the adequacy of the Turing Test itself. It may turn out that machines that are eventually capable of passing the Turing Test clearly have none of these properties. But no matter. The important claim for my purposes is that at some point in the future machines will possess intelligence, self-consciousness and projects that matter to them, as a number of writers hold. I am assuming that an ability to pass the Turing Test will pick out such machines, but if it turns out otherwise, the argument that follows will stand, as long as some machines, perhaps those capable of passing more stringent tests, have these qualities.

Moral standing and personhood

What sort of moral standing should be granted such artificial intelligences? What level of moral concern or regard would we owe to them? Obviously, entities can possess moral standing to different degrees. Most of us would allow that of the various sorts of things that might make moral claims upon us, some of these are capable of sustaining greater claims than others. For instance, we may have a basic level of moral concern for the lower animals, such as fish and crustaceans, a greater concern for mammals such as dogs and elephants, etc, and more still for the higher

primates such as the great apes. It may turn out that intelligent machines should be granted moral standing somewhere along this scale. However in this paper I am concerned to investigate whether machines could achieve moral standing comparable to that we accord normal adult human beings? Could machines be 'moral persons'?

THE 'TURING TRIAGE TEST'

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. 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.[7] 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'. It is my hope that the Turing Triage Test will serve as a focus for discussion of issues surrounding the moral status of artificial intelligences and what would be required for machines to achieve moral standing in the same way that the Turing Test has served to focus attention on the question of whether machines could think and what would be required for them to do so.[8]

Obviously proposing a test for when moral standing has been achieved does not tell us whether or not a machine could pass this test. But it does allow us to think productively about what would be necessary for a machine to pass it; that is, what other sorts of things would need to be true for a machine to achieve moral standing.

THE CASE FOR MORAL STANDING FOR INTELLIGENT MACHINES

Let me begin by observing that according to an influential, perhaps the dominant, account of the nature of personhood, we should expect that future intelligent machines will, sooner or later, be able to pass the Turing Triage Test. A number of philosophers have argued that we should separate our account of the origin of moral concern and of the nature of personhood from the concept of a human being (See, for instance, Singer, 1981; Singer, 1986; Tooley, 1986. Diamond, 1991a, provides a neat paraphrase of this position). Whatever it is that makes human beings morally significant must be something that could conceivably be possessed by other entities. To restrict personhood to human beings is to commit the error of chauvinism or 'speciesism'.

The precise description of qualities required for an entity to be a person or an object of moral concern differ from author to author. However it is generally agreed that a capacity to experience pleasure and pain provides a prima facie case for moral concern and that the grounds for this concern, as well as its proper extent, are greater the more a creature is conscious of itself as existing across time, has its own projects and is capable of reasoning and rationality.

It is a recognised, indeed an intended, consequence of such accounts that they allow that in some cases other entities might have more of these qualities than a given human being. For instance, we might sometimes be obligated to preserve the life of an adult chimpanzee over that of a brain damaged human baby on the grounds that the former has superior cognitive capacities and therefore greater claim to moral regard than the latter. I mention this point to establish that it will not be unprecedented therefore if it turns out that machines sometimes have a better claim to the status of personhood than some human beings.

What is more striking, however, is that it seems as though machines that are plausible candidates for the Turing Triage Test are likely to have a better claim to the status of personhood than **any** human being. If we become capable of manufacturing machines that are apparently capable of self consciousness, reasoning, and investment in personal projects to the **same** extent as a human being, then we will presumably be able to produce machines that are capable of all of these to a much **greater** degree than human beings. There seem to be no reasons to believe that human beings define the upper limit of an ability to do these things. Once we discover how to make machines with such capacities we can simply expand them, perhaps indefinitely. Machines, will after all, not be limited by having a fixed set of capacities available to them due their hardware – as are human beings.

Thus, it is easy to imagine machines that are more intelligent than any human being, more rational, capable of more intricate chains of reasoning, of remembering

and considering more facts and taking into consideration a wider range of arguments. Similarly, intelligent computers may have a greater sense of themselves as entities that endure across time than we do. Their consciousness of self may extend further in both directions than ours; they may have better (more reliable, longer lasting) memories than ours, that allow them to recall exactly what they were thinking at any given moment of their existence; they may have a justified expectation of a vastly longer lifespan than that available to human beings (need intelligent machines fear death due to ordinary circumstances at all?) and so have adopted projects that extend into the distant future. In so far as reasoning capacity, self-consciousness and possession of long term projects is relevant to personhood, future machines are therefore likely to have a greater claim to personhood than do humans.

It might be objected that while I may provide my hypothetical AIs with self consciousness by fiat, it is far from clear that such entities could properly be said to 'suffer'. The ability to experience pleasure and pain might plausibly be held to inhere only in living creatures with nervous systems sufficiently similar to our own. Machine pain can never be anything other than a figure of speech, a rough analogy justified by its usefulness in explaining behaviour (as in, for instance, a case where a robot retreats from a flame that is burning it). Unless machines can be said to suffer they cannot be appropriate objects for moral concern at all.

In fact I have a great deal of sympathy with this objection, as will become clear below. But for the moment I want to outline a popular response to the denial of the

possibility of machine suffering. This response denies that pleasure and pain are states that may only be possessed by living entities and argues that they are properly understood as informational states that can be possessed by any system that behaves in ways suitably analogous to the nervous systems of living creatures regardless of what such systems are made of. Crudely, it is what mental states such as pleasure and pain (and indeed all other cognitive and affective states) **do**, rather than what the mind that experiences them is made of, that makes them what they are.[9] Machines may properly said to possess pleasure and pain states if they have internal states that are 'functionally isomorphic' to similar states in us (Putman, 1975, 291-2).

This argument should, I believe, carry weight with anyone who is prepared to allow that a machine that can pass the Turing Test is intelligent. Presumably intelligent machines have other mental states, such as beliefs and desires, despite the fact that they are made of silicon and metal instead of flesh and blood. Further argument would be required to show why such machines should not be said to suffer or to experience pleasure when they behave in ways appropriate to these states. This is not to say that such arguments could not be made, indeed I will be making them below; however it is to suggest that the onus is on those who would accept the possibility of machine intelligence to explain why the mental life of machines could not include pleasure or pain of the sort that generates moral concern when we witness it in other creatures.

Notice also that if we can imagine a machine with the same capacity to experience pleasure and pain as a human being we should also be able to imagine a machine with a **greater** capacity to experience these things than any human being. For instance, if it possesses circuits that fulfil the same functional role as our nerve endings and pain receptors, we can easily imagine it having **more** of these than we do nerve endings and pain receptors. If it has internal states that map onto our experiences of graduated pleasures or pains then we can imagine it having states that are relevantly analogous to our experiences of great pleasure or great pain, as well as further states that are like these only more so, such that it experiences greater pleasure or greater pain than we do. Again, there seem to be no principled reasons to hold that no entity has a higher capacity for enjoyment or suffering than do human beings.

If a popular philosophical account of the nature of personhood, as consisting in the possession of certain cognitive capacities, is correct, and if we believe what AI enthusiasts say about the likely capacities of future machines, then not only will such machines pass the Turing Triage Test but they are also likely to dominate it. There may be a brief period where machines have only roughly the capacities of human beings and so the choice as to whether to save a human life or to preserve the existence of an intelligent machine constitutes a moral dilemma, but eventually, as the capacities of the machines increase we will always be obligated to save the machine.

I trust this will strike at least some readers as a counter intuitive conclusion, perhaps even a reductio of the arguments considered above. Could it ever really be the case that we were obligated to preserve the existence of a machine, a device of metal and plastic over the life of a member of our own species? For the remainder of this paper I will survey a set of arguments that suggest that it could not.

HUMANS, PERSONS AND AIs

Drawing on the thought of Wittgenstein (1989), a number of philosophers have argued that criterial accounts of personhood of the sort considered above are manifestly inadequate. Personhood is not a matter of having certain capacities or of being able to complete certain tasks.[10] Instead it is a matter of being a creature of a kind such that certain moral and affective responses are appropriately called into existence - and may even be mandatory - in its presence. It is to occupy a certain place in a network of interdependent concepts and moral and affective responses that make up our form of life. This belief that we need to take account of the cluster of concepts and moral and affective responses surrounding our concept of a person derives from a conviction that philosophy needs to pay more attention to the forms of life in which our concepts are embedded. To theorise without due attention to the ways we actually employ our moral language and what we do and do not - and can and cannot – say in it, is to risk losing our way in our investigations. Our concept of a person cannot be adequately captured without paying attention to the ways in which

we behave around and towards people and the various ways in which these differ to our attitudes and behaviour towards non-persons such as animals. Moral emotions such as grief, remorse, sympathy and shame, amongst others, surround and inform our concept of a person.

Because this alternative analysis of personhood links our concept of a person to a wide range of interdependent moral and affective responses, it is possible to begin a challenge to the idea that a machine could pass the Turing Triage Test at a number of points. I will begin my discussion with an analysis of the relationship between personhood and the demands of remorse, grief and sympathy. However this discussion leads quickly to a consideration of the nature of individual personality and the question of the authenticity of the 'suffering' and other internal states of machines. My argument here in turn, rests on observations about the nature of the embodiment of machines.

The concept of a person and the moral emotions

Let us return briefly to reconsider the original triage situation involving a decision as to which of two human lives to save. What makes this situation a moral **dilemma** is the fact that no matter which life we decide to save we have grounds for remorse. A human life, a unique individual life, deserving of the most profound moral respect, has been cut short and this is tragic. It is entirely appropriate that a person required to make such a choice should be haunted by remorse, even in the case when they feel

they could have decided no other way. We would understand if the person required to make this decision was haunted by it for a number of years, even perhaps for their whole life. There may be circumstances, or psychological stories to be told, which explain why an individual did not feel such remorse in a particular case. Perhaps the tragedy was quickly eclipsed by greater tragedy, perhaps the individual concerned has some temporary psychic deficit that prevents them from feeling remorse in this case. However a person who claimed they could not imagine feeling remorse for their decision in the situation would thereby demonstrate a failure to appreciate its nature as a moral dilemma.

Furthermore, it is an integral feature of remorse that it presents itself as a response to the particular individual that we have wronged. As Gaita (1991a, 150-4) has argued, while remorse can only be occasioned by an evil, a transgression of the moral law, it is not directed towards that transgression but towards its victim. The orientation of remorse towards the individual wronged is evidenced in the fact that it is entirely possible that the person experiencing remorse should see the face of the person they have wronged in their dreams, or be tormented by the memory of their voice. The possibility of the intrusion of the personality of their victim into the consciousness of a perpetrator who feels remorse for an evil that they have done is not an extrinsic feature of an ethical response that could be characterised independently. When cases like this occur they are paradigm cases of this moral emotion.

Similarly, it is internal to our sense of the weight of the decision that has to be made that we can imagine that someone should grieve for the individual whose life is lost when medical support is withdrawn. We can also imagine feeling sympathy for their suffering while they await the decision or as they die. Even if we do not feel any such things ourselves, if we cannot conceive of someone doing so, then it seems we do not face a dilemma. Notice also that like remorse, these attitudes are responses to the particular individual whose death or suffering provokes them.

Now let us turn to consider the case when one of the patients in the triage situation has been replaced by an AI. If this is also to be a moral dilemma, it must be appropriate that we might feel remorse no matter which way we made the decision. It must therefore be appropriate that a person might experience remorse for the wrong that the have done the machine in choosing to end its existence. Furthermore, it must be conceivable that this remorse should be such as to haunt a person for years, perhaps even blight their own life. It must also be imaginable that a person should experience grief following the 'death' of the AI, or sympathise with its plight while it awaits a decision, or suffers as power is withdrawn from it.

Is it plausible to believe that someone should be haunted by the evil they have done to a machine? Or that they should feel grief following its death, or sympathy for its suffering. I argue it is not, for reasons that will be expanded upon below. But a first approximation is that machines cannot achieve the sort of individual personality that remorse, grief and sympathy respond to. This is not to claim that machines could not

display unique characteristics such we could distinguish one machine from another. Instead it is to deny that this differentiation could ever establish personality in the richer sense of having a unique inner life of their own. We cannot take seriously a person who claims to feel these moral emotions for a machine because we cannot seriously entertain the idea that machines feel anything at all.

It may seem as though am I making an empirical claim here about our possible responses to machines and a false one at that. Some people clearly do attribute individual personality to machines, as well as emotional states, including suffering. For instance, it is not uncommon to hear people talk of their laptop, or VCR, or even their car, as ‘temperamental’, ‘sulking’, ‘in a bad mood’, or attributing other emotions to these devices. It is less common but still occasionally possible to find people who say that they feel sorry for a machine, or that they experience grief when a machine ‘dies’. Some individuals may even claim that they do feel remorse for wrongs that they believe that they have done to machines. If people already experience such emotions in relation to existing machines with their very limited expressive capacities then how much more likely are they when machines can talk and interact with us.

Yet it is not insignificant here that we feel compelled to enclose the emotional states that people ascribe to machines in inverted commas. We do not **really** believe that they have these emotions. Such descriptions are mere figures of speech or, alternatively, regrettable excesses of sentimentality. Notice that if one really did believe that machines had feelings worthy of moral concern then one could not be

indifferent to whether or not other people recognised these. Indeed one would have to hold that it constituted callousness to fail to do so. Typically, of course, people do not believe this; evidence that they do not themselves really believe what they say. But more importantly, were we to meet someone who held this belief, apparently in all seriousness, I believe we would be forced to conclude that they were misusing the language. The point here is a conceptual rather than an empirical one. It concerns how far it is possible to extend our concepts before they tear loose from the supporting set of responses that give them their meaning.

Embodiment and the 'inner life'

Why should it be so difficult, indeed ultimately impossible, to take seriously the idea that we should feel remorse, grief or sympathy for a machine, or that a machine could be suffering? It is because these responses are only conceivable in relation to creatures which look like us in certain ways. There is a connection between the capacity to engage a certain set of moral responses, including remorse, grief and sympathy, that inform and reflect our sense of the uniqueness of persons that is integral to their moral standing, and possession of a certain sort of physical presence. Crucially, this presence includes possession of a face with expressive eyes and features, and an animate body with the expressive capacities demonstrated by living things (Gaita, 1999, 269).[11] More controversially, it also seems to require that a person be a creature of 'flesh and blood'. Machines are simply not the right sort of things of which to say that they suffer or feel. They lack expressive capacities of the

sort required to ground a recognition of the reality of their inner life. No matter how sophisticated a machine is, it will always remain open to us to doubt the veracity of its purported feelings.

To see this, consider a case where the machine whose feelings we are being called upon to show concern for looks like a filing cabinet with a large number of flashing diodes on the front. This machine has sufficient number of diodes and can flash them in patterns of sufficient complexity so as to demonstrate behaviours that are 'functionally isomorphic' to our pain responses and other cognitive states. The engineer who has designed this machine explains to us that **this** pattern of flashing lights means that the machine is suffering a small pain, **that** one that it is suffering a great pain, this one that it is happy, that one that it is sad, etc. In the light of this information we adjust our behaviour in relation to the machine, in order to minimise its 'suffering'.

Now imagine that the engineer returns to us in a fluster; she has been consulting the wrong manual and has misled us. In fact it is **these** lights which flash when the machine is in pain and **these** lights when it is happy, etc. We should be treating the machine entirely differently. At this point the possibility of radical doubt emerges. How do we know that the engineer has got it right this time? More sinisterly, how do we know that the machine isn't manipulating us by displaying a set of emotions that it is not feeling? How do we know what the machine is **really** feeling?

Once this radical doubt occurs to us, we have no way to resolve the issue. No analysis of the behaviour or structure of a machine will serve to establish that it really feels what it appears to or even that it feels anything at all. There is simply no way of establishing a bridge between the machine's behaviour and any judgements about its purported inner life. It is this unbridgeable gap that opens up between reality and appearance in relation to the thoughts and feelings of machines that explains why we find it impossible to take seriously the thought that machines could have an inner life.

'An attitude towards a soul'

Our doubt about the inner life of machines stands in stark contrast to the knowledge that we possess about the inner lives of the human beings around us. Indeed to call our relation to the internal states of others 'knowledge' is actually a misnomer. It is knowledge only in that it makes no sense to doubt it. We don't even believe that other people have minds, experience pleasure and pain, emotions, etc. No inference is required to reach the conclusion, for example, that someone is in pain when they burn themselves. We simply see it.[12] The fact that such responses are normative for us is evidenced in the way that we question those who do not have them. A person who doubts that others around them have inner lives is not a paragon of rationality who resists an inference that other weaker minds draw without sufficient justification. They are someone who has lost their way in relation to the question entirely. Similarly, to describe these states as 'internal' is misleading. They are not on the

inside of the other person in a way that could be easily contrasted with their outside. They are states of the person and are sometimes visible as such.

Our awareness of the reality of the inner lives of other people is an function of what Peter Winch (1981), following Wittgenstein, calls ‘an attitude towards a soul’. It is a ‘primitive reaction’, a precognitive awareness that is a condition rather than a consequence of our belief that those around us have thoughts and feelings (Winch, 1981, 8; Gaita, 1999, 263-7). Importantly such an attitude is both evidenced in and arises out of a large, complex, and often unconscious set of responses to, and behaviours around, the bodies and faces of other human beings. The fact that we wince when we see another person crack their head, that we can be called into self-consciousness by the gaze of another, that when we bind someone’s wound we look into their face (Gaita, 1999, 266), are all examples of an attitude towards a soul. We cannot help but have such an attitude towards other human beings (Winch, 1981, 11). Conversely, we can not have such an attitude towards a machine.

Androids

Given the role played by the face and the expressive body in creating in us an ‘attitude towards a soul’ towards each other, it might be thought that the question arises as to whether machines with these features could evoke such an attitude in us. There seems to be no reason why AIs could not be provided with expressive faces. Robotics researchers at various labs around the world are already working on faces

for robots in order to facilitate robot/human communication and interaction.[13] Similarly, work on creating humanoid robots is already well advanced. Eventually, perhaps, artificial intelligences will be embodied in androids of the sort made popular by speculative fiction and films such as ‘Blade Runner’, ‘The Terminator’, ‘Alien’, ‘Aliens’ and ‘AI’.

Would we have ‘an attitude towards a soul’ in relation to such androids? It is tempting to allow that we would. After all, by hypothesis they have bodies and faces with the same expressive capacities as those of human beings.[14] Indeed, they may be indistinguishable from human beings. Yet I believe this would be a mistake. To see why, we must return to the discussion above of how a destructive doubt arises in relation to the thought and feeling of machines. It would not, I believe, alter the force of the example if, instead of a box-shaped device with flashing lights, we confronted a bipedal machine with an animatronic ‘face’. Doubt regarding the relation between its internal states and its external appearance and behaviour could still arise and this is sufficient to destroy any attitude towards a soul that might otherwise exist. The artefactual nature of machines means that the question of the real nature of their design is ever present. Thus, while we might be fooled into evincing such an attitude by machines of sufficiently clever construction, if we were to become aware of their nature we would be forced to reassess our attitude towards their supposed thoughts and feelings and deny that they were ever anything but clever simulacra. If I am right in this, then even artificially intelligent androids will fail the Turing Triage Test. Only

creatures of ‘flesh and blood’ with expressive bodies and faces are capable of being the object of an ‘attitude towards a soul’.

Summary

The argument of this final section of the paper has been a complex and difficult one. To recap briefly; I have argued that for a machine to pass the Turing Test it must be capable of being the object of remorse, grief and sympathy, as moral emotions such as these are partially constitutive of our concept of a person. But, I claim, machines are not appropriate objects of these responses because they are incapable of achieving the individual personality towards which they are oriented. We cannot seriously hold that machines have thoughts or feelings or an inner life because a radical doubt inevitably arises as to whether they really feel what they appear to. My argument here in turn draws upon an analysis of the nature of our knowledge of other minds as consisting in ‘an attitude towards a soul’.

A critical reader no doubt may wish to challenge this chain of reasoning at any number of points. I can do no more to defend my account here. However, in closing, I wish again to emphasise the demanding and counter-intuitive nature of the conclusions that are likely to follow from any successful such challenge. For example, that we should feel sympathy for the ‘suffering’ of an entity which has all the expressive capacity of a metal box, and, not least, that we may be obligated to preserve its existence over that of a living human being!

CONCLUSION

The prospect of future artificially intelligent machines raises the question of the moral standing of such creations. Should they be treated as persons? I have offered the 'Turing Triage Test' as a useful device for testing our intuitions in relation to this matter. A popular philosophical account of the nature of persons as beings with a certain set of cognitive capacities leads quickly to the conclusion that not only will future machines pass this test but that they may come to have more claim for moral regard than any human being. However this unpalatable conclusion may also be taken to indicate deep problems with this account of the nature of persons. An alternative account of what it is to be a person, set out most clearly in the work of Raimond Gaita, looks to an interdependent network of moral and affective responses to delineate and give content to our concept of a person. Until AIs are embodied in such a fashion that they can mobilise these responses in us, they will be unable to pass the 'Turing Triage Test'. Unless we could sympathise with the suffering of an AI as we moved to throw the switch that would end its existence, grieve for its 'death', and be haunted by remorse at the thought of the life that we have ended, it would not be reasonable to preserve its existence over that of a human being. It will not be possible for us to properly apply these concepts while artificial intelligences have the character and appearance of machines.

References:

Block, N. (1980), 'Introduction: what is functionalism?' in N. Block, ed., Readings in philosophy of psychology, Vol 1., Cambridge, Mass.: Harvard University Press, pp. 171-184

Churchland, Paul M. (1984), Matter and Consciousness, Cambridge, Mass.: MIT Press

Diamond, C. (1978), 'Eating Meat and Eating People', Philosophy 53, pp. 465-479

Diamond, C. (1991a), 'The Importance of Being Human', in D. Cockburn, ed., Human Beings, Cambridge: Cambridge University Press, pp. 35-62

Diamond, C. (1991b), The Realistic Spirit: Wittgenstein, Philosophy and the Mind, Cambridge, Mass.: MIT Press

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

Floridi, L.L. and Sanders, J.W. (2000), 'Artificial Evil and the Foundation of Computer Ethics', in Deborah Johnson, James Moor & Herman Tavani, ed., Proceedings of Computer Ethics: Philosophical Enquiry 2000, Dartmouth, pp. 142-54

Ford, Kenneth M., Glymour, Clark and Hayes, Patrick J. (1995), ed., Android Epistemology, Cambridge, Mass.: MIT Press

Gaita, R. (1989), 'The Personal in Ethics', in D. Z. Phillips and P. Winch, ed., Wittgenstein: Attention to Particulars, pp. 124-150

Gaita, R. (1990), 'Ethical Individuality', in R. Gaita, ed., Value and Understanding, London: Routledge, pp. 118-148

Gaita, R. (1991a), Good and Evil: An Absolute Conception, London: MacMillan

Gaita, R. (1991b), 'Language and Conversation: Wittgenstein's Builders', in A. P. Griffiths, ed., Wittgenstein Centenary Essays, Cambridge: Cambridge University Press, pp. 101-15

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

Jackson, Frank and Pettit, Philip (1988), 'Functionalism and Broad Content', Mind 97, pp. 381-400

Kurzweil, Ray (1992), The Age of Intelligent Machines, Cambridge, Mass.: MIT Press

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

Levinas, E. (1989), 'Ethics as First Philosophy', in S. Hand, ed., The Levinas Reader, Oxford: Basil Blackwell, pp 75-87

Lingis, A. (1994), The Community of Those Who Have Nothing In Common, Bloomington and Indianapolis: Indiana University Press

Menzel, P. and D'Aluisio, F. (2000), Robo Sapiens: Evolution of a New Species, Cambridge, Mass.: The MIT Press

Moravec, Hans (1988), Mind Children: The future of robot and human intelligence, Cambridge, Mass.: Harvard University Press

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

Putnam, H. (1975), 'Philosophy and our mental life', in Mind, Language, and Reality, Cambridge: Cambridge University Press, pp. 291-303

Saygin, A.P., Cicekli, L. and Akman, V. (2000), 'Turing Test: 50 Years Later', Minds and Machines 10(4), pp. 463-518

Singer, Peter (1981), The Expanding Circle: Ethics and Sociobiology, New York: Farrar, Straus & Giroux

Singer, Peter (1986), 'All Animals are Equal', in P. Singer, ed., Applied Ethics, Oxford: Oxford University Press, pp. 214-228

Simons, Geoff (1992), Robots: The Quest for Living Machines, London: Cassell

Tooley, M. (1986), 'Abortion and Infanticide', in P. Singer, ed., Applied Ethics, Oxford: Oxford University Press, pp. 57-86

Turing, Alan (1950), 'Computing machinery and intelligence', Mind 59, pp. 433-60

Winch, Peter (1981), 'Eine Einstellung zur Seele', Proceedings of the Aristotelian Society

Wittgenstein, L. (1989), Philosophical Investigations, Oxford: Basil Blackwell

---

[1] This will mark the beginning of a new field that might be called 'Android Ethics', to go alongside 'Android Epistemology'. Cf Ford, Glymour and Hayes (1995). The birth of a new field of 'Android Ethics' is also heralded in Floridi and Sanders (2000).

[2] Although it is worth noting that even the more modest systems designed by 'weak AI' researchers may have some claim to moral regard and raise some of the issues with which I am concerned here. These may be usefully illuminated by considering the limit case of whether intelligent computers might achieve the moral status of persons.

[3] In fact this is a simplification of the original Turing Test, which required that a machine be as good as a man at pretending to be a woman. That is, the task of the computer is to be equally as good as a human being at an imitation game involving gender. The role played by gender in the original formulation of the Turing Test is usually neglected in later discussion of the test (Saygin, Cicekli and Akman, 2000).

[4] For a recent survey of the literature surrounding the Turing Test, see Saygin, Cicekli and Akman (2000).

[5] Although it must be noted that the conclusion of my paper suggests that passing the Turing Test will be much more difficult for machines than is currently recognised. Indeed it may well be impossible. My assumption that machines will pass the test is a working hypothesis for the purposes of the argument of the paper.

[6] It might be objected that a machine's ability to report on its internal state does not establish that it **really** has these features. As will become obvious later, I agree with this objection. But at first sight it

does not seem to distinguish between the case of a machine's consciousness or desires and its intelligence. If the Turing Test is an adequate test for the presence of intelligence then further argument is required to show why it will not be adequate for these other qualities.

[7] Let me also stipulate that neither of the available courses of action will lead to any further loss of life. The remaining patient is not a doctor or scientist whose advice is urgently needed. The hospital will be able to treat its patients properly without the AI in the short period before an alternative source of diagnostic advice can be found. In this decision, the only relevant consideration is the moral status of the two claimants in front of you.

[8] The Turing Triage Test may also be relevant to the question of the moral status of cyborgs – human/machine hybrids - although I will not be able to examine the matter here.

[9] Functionalism avoids the problems of behaviourism by allowing that the nature and identity of these states is determined not only by their relation to the external behaviour of the system but also to other such internal states. See Block (1980), Churchland (1984), Putman (1975), Jackson and Petit (1988).

[10] My discussion below largely follows arguments presented in Gaita (1989, 1990, 1991a), and further refined in Gaita (1999). A central claim in Gaita's argument was previously developed in an important and difficult paper by Peter Winch (1981). Cora Diamond (1978, 1991a, 1991b) covers related territory in a number of discussions of the ethics of our treatment of animals. Gaita and Winch draw heavily on the discussion of the nature of pain and of pain attributions in Wittgenstein (1989).

[11] The role played by the human face at the very foundation of the nature of our ethical relationship with others is also argued for by Levinas (1989), Lingis (1994).

[12] Of course there **are** cases when we do wonder whether or not the emotion or pain someone appears to feel is real. We may, for instance, believe that they are acting a part or trying to deceive us. But here the possibility of such questioning is established by the certainty we have in ordinary cases. While the question of the veracity of the feelings displayed by another person may arise in particular cases, it never seriously occurs to us to doubt that this individual has thoughts and emotions, let alone that people in general have an internal life (Gaita, 1999, 263-7).

[13] For example, researchers at the MIT Artificial Intelligence Laboratory have developed 'Kismet', a robot that is designed to respond to human facial expressions and can express its own 'emotions' through its own caricature like face. For an introduction to Kismet and to other contemporary robot research in this area see Menzel and D'Aluisio (2000).

[14] It is worth pausing to note how demanding this assumption really is. Could any combination of metal and plastic achieve the near infinite expressive power of that is possessed by human flesh? Could it allow the empathic awareness of another's emotions that exists, even in silence, between lovers and friends? Our justified cynicism about such a possibility may go a long way towards explaining our reluctance to believe that even androids might possess an inner life.