Ethics and Information Technology (2004) 6: 203–213
DOI 10.1007/s10676-004-6491-2

The Turing Triage Test

Robert Sparrow
School of Philosophy and Bioethics, Faculty of Arts, Monash University, Victoria 3800, Australia; Centre for Applied Philosophy and Public Ethics, University of Melbourne, Australia
E-mail: Robert.Sparrow@arts.monash.edu.au

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. Inspired by 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 that 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.

Key words: artificial intelligence, computers, ethics, Turing Test, person, embodiment

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 21st century – and according to some, well before this – machines will be conscious, intelligent, entities with capacities exceeding our own (Moravec, 1988; Kurzweil, 1992, 1999; Simons, 1992; Moravec, 1998; Dyson, 1997).

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.

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

In this paper I propose a test to determine when we have reached that point. Inspired by 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 that 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 of 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

[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.