How to Detect an Android

Antoni Diller
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK

A.R.Diller@cs.bham.ac.uk

http:/www.cs.bham.ac.uk/∼ard

September 2011

Abstract

A number of prominent researchers in Artificial Intelligence look forward to a time when androids will co-exist with humans. In such a world the question 'Can machines think?' will be answered in the affirmative, if it is asked at all. A far more pressing problem will be whether androids can be differentiated from humans in some way. It is argued that an undetectable android cannot be manufactured and a method is presented for distinguishing between androids and humans.

1 Introduction

Turing [22] famously devised a test, the imitation game, whose purpose was to investigate the question whether human intellectual ability was significantly different from that of a machine. The nature of the test was influenced by the technology of the day. An interrogator and two test subjects, one male and one female, occupy different rooms. (The interrogator can be either a man or a woman.) They communicate by means of a teleprinter. The interrogator has to determine which of the subjects is male and which female. He does this by asking both of them a series of questions. The man is instructed to try and fool the interrogator into making an incorrect identification, whereas the woman is told to aid the interrogator in his or her task. The game is played in this form several times. The male subject is then replaced by a computer. Presumably, the interrogator knows this has taken place. (Unfortunately, Turing's account of the imitation game is quite sketchy.) It is now the computer's task to fool the interrogator into making the wrong identification. Again, the game is played in this form several times. Turing [22, p. 434] replaces the question 'Can machines think?' with: 'Will the interrogator decide wrongly as often when the game is played like this [that is, between a machine and a woman] as he [or she] does when the game is played between a man and a woman?' (Turing thought a time would come when he or she would.) This question presupposes the game is played several times in both versions. It is not clear, however, if the participants are different each time. I think they would have to be for the procedure to make sense, but as the unearthing of Turing's precise meaning is not my main concern here, I will not pursue the matter.

Inspired by Turing's work, Hugh Loebner created the Loebner Prize in the early 1990s. He offered a sum of $100,000 and a gold medal to the first person who could devise a program that could fool ten judges into thinking it was human on the basis of a conversation, about anything, lasting three hours. This prize has still to be won. A prize for the most human-like program is, however, awarded each year. Programs in this competition are only expected to converse about a single topic for just a few minutes. Reading through transcripts of entrants' dialogues shows that the programs are becoming increasingly sophisticated and that the gold medal may, indeed, be won one day. (These transcripts, and further information about the prize, are available on the `www.loebner.net` website.)

Technology has improved dramatically since Turing's day and many workers in Artificial Intelligence (AI) and Robotics today are far more ambitious than he ever was. They would agree with Brooks and his colleagues when they write: 'Building an android, an autonomous robot with humanoid form and human-like abilities, has been both a recurring theme in science fiction and a "Holy Grail" for the Artificial Intelligence community' [3, p. 52]. Considerable progress has been made in achieving the goal of manufacturing an android and one day it is likely that very life-like machines will co-exist with humans. In such a world the issue will no longer be whether machines can think, but how to distinguish between humans and androids.

2 New Questions

Is it possible to construct a robot that looks like a human being and which can live in human society and be taken for a human being? Is it possible for a machine made out of mechanical and electronic components to behave in exactly the same way as a human? Some scientists working in AI and Robotics believe that it is. Brooks thinks that it is just a matter of time before an android is built that can pass for human [2, p. 209]. Moravec believes this will happen by 2040 [19, p. 88]. In this paper, I argue that it is impossible for human beings to design and build an undetectable android. In particular, I show that human emotion-related behaviour will always be distinguishable from android emotion-related behaviour. I focus on emotion-related behaviour in order to make my reasoning easier to follow. With suitable changes, however, the argument could be used to show that any clearly distinguishable type of human behaviour could never be perfectly replicated by an android. After showing that androids will always be detectable, I present a method that could be used to identify androids once they start living amongst us.

3 Designing Androids

Before any complex mechanism can be built, it first has to be carefully designed. The design team has to make use of a large number of theories. This applies as much to the design of an android as it does, say, to that of an aeroplane. When crafting an aeroplane, the design team needs experts in many fields, including aerodynamics, ballistics, chemistry, mechanics, metallurgy and structural engineering. Similarly, when fashioning an android, the design team will comprise people with expertise in several disciplines. They will require knowledge of theories concerning many different things. Some of these will be physical theories relating to the android's mechanical components. The android's limbs, for example, may be moved using hydraulic pumps and so some members of the design team will need to have a thorough knowledge of hydraulics. Some of the theories, however, will relate to the android's social behaviour. Knowledge of physics will enable the design team to produce an android that can shake hands, say, but not one that knows when it is socially acceptable to do so. I am interested in those theories that relate to the android's social behaviour and intellectual abilities. Some psychologists use the term 'social brain' to refer to that aspect of a human being responsible for producing emotions and emotion-related behaviour. This is a convenient label, but by using it I do not wish to suggest that I believe there is an identifiable region of the physical brain that is responsible for everything to do with emotions in human beings. In this paper, I am particularly interested in the problem of designing an android's social brain.

It should be noted that, using a distinction introduced by Dennett [5, p. 194], the android would *instantiate* the theories used in its design and not merely *incorporate* them. A computer simulation of a hurricane, say, incorporates a theory of hurricane behaviour. It is a computer 'program, which, when you feed in *descriptions* of new meteorological conditions, gives you back *descriptions* of subsequent hurricane' behaviour (p. 191); it does not produce such behaviour. The android I am considering, however, would instantiate various theories, because its behaviour would be generated by those theories; they would not produce descriptions of human behaviour.

In order to live in a human society and interact meaningfully with human beings an android must be able to display and recognise human emotions. Currently, our understanding of the nature of emotion is lamentable. There are many different theories of emotion competing for our attention. For example, Strongman [21], in chapter 2 of his book *The Psychology of Emotion*, discusses about thirty different theories, but he does not claim to have mentioned every theory there is. He admits [21, p. 13]: 'To describe all theories of emotion would necessitate a book in itself, so the number has been restricted.' The fact that there are so many competing and incompatible theories strongly suggests that none of them are very good, for the existence of a good theory, or a small number of good theories, would quickly lead to its competitors being consigned to the dustbin of history. So, using an existing theory of emotion in the design of an android's social brain is highly unlikely to result in the android exhibiting emotion-related behaviour which is indistinguishable from that of a human. However, even if our understanding of emotion improved dramatically, that would not increase the chances of making an undetectable android, because, no matter how good our theories of emotion become, they will all always be false. This follows from the fact that all scientific theories are false. (It should be noted that this is very different from saying that they are all useless and worthless, as I explain below.) This claim is an essential premise in my argument to show that there cannot be an undetectable android. To many people this claim seems absurd when they first come across it, but a little investigation shows it is much more widely accepted than you would initially think. Lakatos, for example, liked to say that all scientific theories are born refuted, live in an 'ocean of anomalies' and die refuted [16, pp. 5, 126, 128 and 147]. Furthermore, there is much discussion, by philosophers and historians of science, of an argument known as the *pessimistic induction* (or *meta-induction*). This starts from the observation that the history of science is littered with discredited theories. These include the theory of spontaneous generation, the caloric theory of heat, the theory of global cooling (widely accepted in the 1970s), Aristotelian mechanics, Ptolemaic astronomy, Kepler's celestial mechanics, geological catastrophism, the medical theory of humours, the effluvial theory of static electricity, the theory that the chemical atom is indivisible and the phlogiston theory. This list could easily be extended with many more examples. Newton-Smith [20, p. 183] presents the pessimistic induction as follows:

> Past theories have turned out to be false, and since there is no good reason to make an exception in favour of our currently most cherished theories, we ought to conclude that all theories which have been or will be propounded are strictly speaking false.

He even goes so far as to say [20, p. 14]: 'Indeed the evidence might even be held to support the conclusion that no theory that will ever be discovered by the human race is strictly speaking true.' (In order to avoid paradox it is important to emphasise that the conclusion of the pessimistic induction is not itself a scientific theory of the same sort or level as the theories it applies to. Recent work on the pessimistic induction includes papers by Hobbs [15], Lewis [18], Lange [17] and Busch [4].)

Inductive arguments are not deductively valid and so it is possible for their conclusions to be false even when all their premises are true. They are, however, often useful as heuristics. In science, for example, they are frequently helpful in suggesting universal theories that the scientific community then tries to refute. The inability to falsify a theory increases our confidence in its verisimilitude. Why theories are added to or removed from the body of generally-accepted scientific knowledge is a complex matter which is still being studied by philosophers of science. The difficulty of the issue was revealed by the pioneering work of Kuhn, Lakatos, Feyerabend and Laudan. The main problem is that counter-examples can be found to the various rational criteria of acceptability that have been put forward. (Some of this research is admirably summarised in chapter 4 of Bechtel's book *The Philosophy of Science* [1], written for cognitive scientists.)

The considerations presented show that the conclusion of the pessimistic induction has not been conclusively established, but the pessimistic induction does provide strong *prima facie* evidence in its favour. Much more could be said about the view that all scientific theories are false, but my main concern is not to present an exhaustive account of the debate that the pessimistic induction has given rise to. I will just say, however, that the debate is exclusively about how good an argument the pessimistic induction is. No one, as far as I know, tries to falsify its conclusion by presenting a true and completely accurate scientific theory which is universally acknowledged as such.

If the reader is undecided about the correctness of the view that all scientific theories are false, then he or she may still be interested in how I derive the claim that an undetectable android cannot be made from this position, together with some other considerations, and especially in the following conditional sentence, which is established if my reasoning is correct: If all scientific theories are false, then an undetectable android cannot be made. Someone who is in two minds about the claim that all scientific theories are false is still likely to find the contrapositive of the previous conditional sentence downright counter-intuitive: If an undetectable android can be made, then some scientific theories are true.

Although many people think that is is reasonable to believe that all scientific theories are false, this position does not have disastrous consequences for science or technology. Just because a theory is false it does not follow that it is useless. Newton's theory, for example, is false, but it was used by NASA scientists in order to plot the trajectories of rockets sent to the moon and those rockets reached their destination. It is also regularly used by scientists in their preparations to launch satellites. Although Newton's theory has been falsified, it is still a good approximation to the truth. Furthermore, some theories which we now regard as radically mistaken, such as the medical theory of humours and the eighteenth-century optical aether theory, were incredibly successful and useful in their day.

Returning to the question of designing the social brain of an android, let T be the theory of emotion used by the design team in order to develop and build the android's social brain C. The android's emotion-related behaviour is produced by C. Using Dennett's terminology, C instantiates T. Theory T is a model of human emotion-related behaviour, a conjecture as to how such behaviour is produced by a human's social brain. In these circumstances it would be possible for human behaviour and emotional response to falsify theory T, but it would not be possible for android behaviour and emotional response to falsify T. As a theory of human emotion, T is a falsifiable, empirical theory, whereas as a theory of android emotion, produced by component C, it is an unfalsifiable, non-empirical theory. It would be possible for theory T to be falsified by experiments involving human beings, but it would be impossible for android behaviour to falsify the theory, as that behaviour is produced by T. Because all scientific theories are false, we know that T is false and some human behaviour does actually falsify it. This means that the difference between human emotional response and android emotional response can be established by any human behaviour that falsifies theory T, because an android could not exhibit such behaviour. Thus, there is a way of detecting the presence of androids in human society.

An example may clarify the above argument. This example has been simplified, but it still accurately presents the issues involved. Consider a situation in which an android has been equipped with a cognitive theory of emotion as developed, for example, by Dryden [14]. The specific emotion that I shall discuss is anxiety. In such a theory people feel anxious and exhibit anxiety-related behaviour if they have a cluster of irrational beliefs about some imminent, personally significant event. For example, imagine someone facing a job interview and holding the irrational beliefs: 'I must get this job', 'It would be terrible if I failed to get this job', 'I couldn't stand not getting this job' and 'If I fail to get this job, that proves I'm worthless and that I'll never get a decent job'. According to the cognitive theory such a person would feel extremely anxious and exhibit some anxiety-related symptoms. For the sake of argument, I take these symptoms to be disturbed sleep, dryness in the mouth, trembling and sweating. An android equipped with this theory of emotion, holding these irrational beliefs and facing a job interview would exhibit the anxiety-related symptoms listed above. Moreover, whenever it was going for a job interview having these irrational beliefs, it would exhibit the same symptoms. In the case of a human being, however, it is not logically impossible to conceive of a person holding these irrational beliefs, while waiting for a job interview, and yet not exhibiting some of these symptoms. In fact, because of the plausibility of the claim that all scientific theories are false, we can be confident that the cognitive theory of emotion, no matter how useful it is in psychotherapy, is false and some human behaviour does actually falsify it.

For the sake of argument, let us assume that a person with the above irrational beliefs facing a job interview does not exhibit any trembling. An android in whom the cognitive theory was instantiated would tremble in these circumstances. Thus, the presence of trembling in an entity looking like a human being would alert us to the fact that we were dealing with an android.

Perhaps an analogy will make my reasoning clearer. An orrery is a clockwork model of the solar system. An orrery could be built to illustrate Ptolemy's celestial mechanics. The Sun and planets would move around the Earth in orbits produced by combining several circular motions. Another orrery could be made to illustrate Copernicus's theory in which the orbits of the planets are again obtained by combining several circular motions, but now the Sun is at the centre of the system. Yet another orrery could be fashioned to illustrate Kepler's celestial mechanics. In this the planets would move in ellipses around the Sun which would lie at one of the foci of each of these ellipses. In reality, an orrery could not be built which is an exact scale model of the solar system, because the highest common factor of the mean distances of the planets from one another is minute in comparison with the mean distance of the furthest planet from the Sun. A computer simulation could, however, be fashioned. Imagine such a simulation built to illustrate Kepler's celestial mechanics. The behaviour of the simulation could not deviate from that described by Kepler's theory. No matter how many readings we took of the simulation they would always be in conformity with that theory. It would be impossible for any such readings to falsify Kepler's theory. We know that the behaviour of our solar system is different from what we would expect on the basis of Kepler's theory. The behaviour of our solar system that falsified Kepler's celestial mechanics could not be produced by the computer simulation.

In this analogy, the computer simulation (or orrery) corresponds to an android and the behaviour of the simulation corresponds to the android's emotion-related behaviour, the real solar system corresponds to a human being and its behaviour corresponds to human emotion-related behaviour. Kepler's celestial mechanics corresponds to the theory of emotion used to produce the android's emotion-related behaviour. In certain circumstances, there are discrepancies between android and human behaviour, just as there are discrepancies between the behaviour of the simulation and the behaviour of the solar system. In the case of emotion-related behaviour these discrepancies allow us to differentiate between androids and humans.

4 The Method of Detection

The above argument shows that there must be a discernible difference between human and android emotion-related behaviour. On the basis of these considerations a method can be devised to ascertain whether a test subject is an android or a human. Let us assume that we suspect our subject to be an android manufactured by a particular corporation. We first have to find out which theory of emotion T scientists working for that corporation used in designing their androids. This theory is instantiated in the social brains of all their androids. Then, we have to find out what human behaviour B falsifies T. After that, we have to put the test subject in a situation where a human would exhibit behaviour B. If the subject exhibits B, it is human. If the subject fails to exhibit B, it is an android. As behaviour B falsifies T, it could not be exhibited by the android.

It is possible that different design teams, working for other companies, might use various theories of emotion when designing androids. It might, therefore, be necessary to use the above method several times in order to find out if a test subject is an android. It might be necessary to use the above method as many times as there are theories of emotion that have been instantiated in different ranges of android. Furthermore, as theories of emotion get more sophisticated the behaviour B that we need to discover to detect androids will get harder to find, but so long as we deal with scientific and empirical theories of emotion we can be confident that such behaviour exists.

5 Conclusion

Unlike some people, I am not frightened by the prospect of intelligent, autonomous androids living side-by-side with humans. Furthermore, I believe this will happen one day. I am irritated, however, by people who say that this will happen in the near future, maybe in the next fifty years or so. This is because there are certain human intellectual abilities that hardly anybody is trying to mechanise and, until they are incorporated in androids, humanoid robots will have no chance of competing intellectually with humans. For example, the capacity to acquire information from testimony, that is, from what others say and have written, is crucial to our lives in society. An android would also need to have this aptitude in order to live, work and interact meaningfully with humans, yet virtually no one in AI and Robotics is studying how to give androids this ability. Not until we can give androids this aptitude will they have a chance of being truly intelligent [9, 12, 13]. I am, however, neither a Luddite nor a Jeremiah. I have been studying testimony from an AI perspective in order to work out how to equip androids with the capacity to learn form other people's assertions [7, 8, 10, 11, 13]. It is important that androids should have this ability. Be that as it may, in this paper my purpose has been different. I have argued that it will always be possible to devise a method to tell whether a human-looking being is really human or whether it is an android. This is not meant to deter people from trying to make ever more life-like robots. I want people to continue producing ever more sophisticated androids. It is, however, the fact that they are designed and built by fallible human beings that makes them detectable.

Acknowledgements

A preliminary version of the argument presented here, written for a general readership, appeared in a simplified form in the popular magazine *Philosophy Now* [6]. I have also expounded it in several lectures, talks and seminars over the years; I am grateful for the feedback that I have received.

References

[1] William Bechtel. *Philosophy of Science: An Overview for Cognitive Science*. Tutorial Essays in Cognitive Science (advisory editors: Donald A. Norman and Andrew Ortony). Lawrence Erlbaum Associates, Hillsdale (New Jersey), 1988.

[2] Rodney A. Brooks. *Robot: The Future of Flesh and Machines*. Allen Lane: The Penguin Group, London, 2002.

[3] Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. The Cog project: Building a humanoid robot. In C. Nehaniv, editor, *Computation for Metaphors, Analogy, and Agents*, volume 1562 of *Lecture Notes in Computer Science*, pages 52–87, Heidelberg, 1999. Springer-Verlag.

[4] Jacob Busch. Entity realism meets the pessimistic meta-induction argument. *Sats: Nordic Journal of Philosophy*, 7(2):106–126, 2006.

[5] Daniel C. Dennett. Why you can't make a computer that feels pain. In *Brainstorms: Philosophical Essays on Mind and Psychology*, chapter 11, pages 190–229. Harvester Press, Brighton, 1981.

[6] Antoni Diller. Detecting androids. *Philosophy Now*, (25):26–28, Winter 1999/2000.

[7] Antoni Diller. Acquiring information from books. In Max Bramer, Alun Preece, and Frans Coenen, editors, *Research and Development in Intelligent Systems XVII: Proceedings of ES2000, the Twentieth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2000*, pages 337–348. Springer, London, 2001.

[8] Antoni Diller. A model of assertion evaluation. Cognitive Science Research Papers CSRP–02–11, School of Computer Science, University of Birmingham, November 2002.

[9] Antoni Diller. Designing androids. *Philosophy Now*, (42):28–31, July/August 2003.

[10] Antoni Diller. Modelling assertion evaluation. *AISB Quarterly*, (114):4, Autumn 2003.

[11] Antoni Diller. Assessing information heard on the radio. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining: Proceedings of the International IIS:IIPWM'05 Conference held in Gdańsk, Poland, June 13–16, 2005*, Advances in Soft Computing, pages 426–430. Springer, Berlin, 2005.

[12] Antoni Diller. How empiricism distorts AI and Robotics. In M. H. Hamza, editor, *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2005)*, pages 339–343. ACTA Press, Anaheim, Calgary, Zurich, 2005.

[13] Antoni Diller. Why AI and Robotics are going nowhere fast. In Jordi Vallverdú, editor, *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles*, pages 328–343. Information Science Reference, Hershey (PA), 2010.

[14] Windy Dryden. *Invitation to Rational-emotive Psychology*. Invitations to Psychology. Whurr, London, 1994.

[15] Jesse Hobbs. A limited defence of the pessimistic induction. *British Journal for the Philosophy of Science*, 45(1):171–191, 1994.

[16] Imre Lakatos. *The Methodology of Scientific Research Programmes: Philosophical Papers, volume 1*. Cambridge University Press, Cambridge, 1978. Edited by John Worrall and Gregory Currie.

[17] Marc Lange. Baseball, pessimistic inductions and the turnover fallacy. *Analysis*, 62(4):281–285, October 2002.

[18] Peter J. Lewis. Why the pessimistic induction is a fallacy. *Synthese*, 129:371–380, 2001.

[19] Hans Moravec. Rise of the robots. *Scientific American*, 281(6):86–93, December 1999.

[20] W. H. Newton-Smith. *The Rationality of Science*. International Library of Philosophy, editor: Ted Honderich. Routledge & Kegan Paul, Boston, London and Henley, 1981.

[21] K. T. Strongman. *The Psychology of Emotion*. John Wiley & Sons, Chichester, second edition, 1978.

[22] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, October 1950.