CHAPTER 12

Intelligent Machinery, A Heretical Theory (*c.1951*)

Alan Turing

Introduction

Jack Copeland

The '51 Society

Turing gave the presentation 'Intelligent Machinery, A Heretical Theory' on a radio discussion programme called *The '51 Society*. Named after the year in which the programme first went to air, *The '51 Society* was produced by the BBC Home Service at their Manchester studio and ran for several years.[1] A presentation by the week's guest would be followed by a panel discussion. Regulars on the panel included Max Newman, Professor of Mathematics at Manchester, the philosopher Michael Polanyi, then Professor of Social Studies at Manchester, and the mathematician Peter Hilton, a younger member of Newman's department at Manchester who had worked with Turing and Newman at Bletchley Park.

Machine Learning

Turing's target in 'Intelligent Machinery, A Heretical Theory' is the claim that 'You cannot make a machine to think for you' (p. 472). A common theme in his writing is that if a machine is to be intelligent, then it will need to 'learn by experience' (probably with some pre-selection, by an external educator, of the experiences to which the machine will be subjected). The present article continues the discussion of machine learning begun in Chapters 10 and 11. Turing remarks that the 'human analogy alone' suggests that a process of education 'would in practice be an essential to the production of a reasonably intelligent machine within a reasonably short space of time' (p. 473). He emphasizes the

[1] Peter Hilton in interview with Copeland (June 2001).

point, also made in Chapter 11, that one might 'start from a comparatively simple machine, and, by subjecting it to a suitable range of "experience" transform it into one which was more elaborate, and was able to deal with a far greater range of contingencies' (p. 473).

Turing goes on to give some indication of how learning might be accomplished, introducing the idea of a machine's building up what he calls 'indexes of experiences' (p. 474). (This idea is not mentioned elsewhere in his writings.) An example of an index of experiences is a list (ordered in some way) of situations in which the machine has found itself, coupled with the action that was taken, and the outcome, good or bad. The situations are described in terms of features. Faced with a choice as to what to do next, the machine looks up features of its present situation in whatever indexes it has. If this procedure affords more than one candidate action, the machine selects between them by means of some rule, possibly itself learned through experience. Turing very reasonably grounds his belief that comparatively crude selection-rules will lead to satisfactory behaviour in the fact that engineering problems are regularly solved by 'the crudest rule of thumb procedure . . . e.g. whether a function increases or decreases with one of its variables' (p. 474).

In response to the problem of how the educator is to indicate to the machine whether a situation or outcome is a 'favourable' one or not, Turing returns to the possibility of incorporating two 'keys' in the machine, which can be manipulated by the educator, and which represent 'pleasure' and 'pain' (p. 474). This is an idea that Turing discusses more fully in Chapter 10, where he considers adding two input lines to a (modified) Turing machine, the pleasure (or reward) line and the pain (or punishment) line. He calls the result a 'P-type machine' ('P' standing for 'pleasure–pain').[2]

Random Elements

Turing ends his discussion of machine learning with the suggestion that a 'random element' be incorporated in the machine (p. 475). This would, as he says, result in the behaviour of the machine being by no means completely determined by the experiences to which it was subjected (p. 475). The idea that a random element be included in a learning machine appears elsewhere in Turing's discussions of machine intelligence. In Chapter 11 he says: 'A random element is rather useful when . . . searching for a solution of some problem' (p. 463). He gives this example:

[2] A detailed description of Turing's P-type machines is given in B. J. Copeland and D. Proudfoot, 'On Alan Turing's Anticipation of Connectionism', *Synthese*, 108 (1996), 361–77 (reprinted in R. Chrisley (ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, ii: *Symbolic AI* (London: Routledge, 2000)).

Suppose for instance we wanted to find a number between 50 and 200 which was equal to the square of the sum of its digits, we might start at 51 then try 52 and so on until we got a number that worked. Alternatively we might choose numbers at random until we got a good one.

Turing continues (p. 463):

The systematic method has the disadvantage that there may be an enormous block without any solutions in the region which has to be investigated first. Now the learning process may be regarded as a search for a form of behaviour which will satisfy the teacher (or some other criterion). Since there is probably a very large number of satisfactory solutions the random method seems to be better than the systematic. It should be noticed that it is used in the analogous process of evolution.

Turing's discussion of 'pleasure–pain systems' in Chapter 10 also mentions randomness (p. 425):

I will use this term ['pleasure–pain' system] to mean an unorganised machine of the following general character: The configurations of the machine are described by two expressions, which we may call the character-expression and the situation-expression. The character and situation at any moment, together with the input signals, determine the character and situation at the next moment. The character may be subject to some random variation. Pleasure interference has a tendency to fix the character i.e. towards preventing it changing, whereas pain stimuli tend to disrupt the character, causing features which had become fixed to change, or to become again subject to random variation.

The Mathematical Objection

In what are some of the most interesting remarks in 'Intelligent Machinery, A Heretical Theory', Turing sketches and rebuts an argument against the possibility of computing machines emulating the full intelligence of human beings. The objection is stated as follows in Chapter 10 (pp. 410–11):

Recently the theorem of Gödel and related results . . . have shown that if one tries to use machines for such purposes as determining the truth or falsity of mathematical theorems and one is not willing to tolerate an occasional wrong result, then any given machine will in some cases be unable to give an answer at all. On the other hand the human intelligence seems to be able to find methods of ever-increasing power for dealing with such problems 'transcending' the methods available to machines.

In Chapter 11 he terms this the 'Mathematical Objection' (p. 450).

As Turing notes, the 'related results' include what he himself proved in 'On Computable Numbers'. The import of the satisfactoriness problem (explained in 'Computable Numbers: A Guide') is that no Turing machine can correctly determine the truth or falsity of each statement of the form 'such-and-such

Turing machine is circle-free'. Whichever Turing machine one chooses to ask, there will be statements of this form for which the chosen machine either gives no answer or gives the wrong answer (compare Chapter 11, pp. 450–1). (In Chapter 3, Turing extends this result to his oracle machines: no oracle machine can correctly determine the truth or falsity of each statement of the form 'such-and-such oracle machine is circle-free' (pp. 156–7).)

Post formulated a version of the Mathematical Objection as early as 1921.[3] However, the objection has become known over the years as the 'Gödel argument'. In 1961, in a famous article, the philosopher John Lucas claimed the Gödel argument establishes that 'mechanism'—which Lucas characterizes as the view that 'minds [can] be explained as machines'—is false.[4] More recently, the mathematical physicist Roger Penrose has endorsed a version of the Gödel argument.[5]

Lucas was happy to assert, on the basis of the Mathematical Objection, that 'no scientific enquiry can ever exhaust the . . . human mind'.[6] Not many who admire the explanatory power of science would be happy to endorse this conclusion. Penrose himself appears to hold that the mind can be explained in ultimately physical terms. However, it is difficult to say what scientific conception of the mind could be available to someone who endorses the Mathematical Objection. This is because the objection, if sound, could be used equally well to support the conclusion, not only that the mind is not a Turing machine, but also that it is not any one of a very broad range of machines (which includes the oracle machines). Given the enormous diversity of types of machine in this range, it is an open question whether there is any scientific conception of the mind that the Mathematical Objection (if sound) would not rule out.[7]

Penrose acknowledges that the objection applies not only to the view that the mind is equivalent to a Turing machine but 'much more generally', saying: 'No doubt there are readers who believe that the last vestige of credibility of my [version of the Gödel] argument has disappeared at this stage! I certainly should not blame any reader for feeling this way.'[8]

So far, however, Penrose has not made it clear what scientific conception of the mind can remain for one who endorses the argument, remarking only that, since

[3] E. L. Post, 'Absolutely Unsolvable Problems and Relatively Undecidable Propositions: Account of an Anticipation', in M. Davis (ed.), *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions* (New York: Raven, 1965), 417; see also 423.

[4] J. R. Lucas, 'Minds, Machines and Gödel', *Philosophy*, 36 (1961), 112–27 (112).

[5] See his *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1989); 'Précis of *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*', *Behavioral and Brain Sciences*, 13 (1990), 643–55 and 692–705; *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford: Oxford University Press, 1994); 'Beyond the Doubting of a Shadow', *Psyche*, 2/23 (1996).

[6] Lucas, 'Minds, Machines and Gödel', 127.

[7] See B. J. Copeland, 'Turing's O-machines, Penrose, Searle, and the Brain', *Analysis*, 58 (1998), 128–38.

[8] Penrose, 'Beyond the Doubting of a Shadow', section 3.10, and *Shadows of the Mind*, 381.

the argument 'can be applied in very general circumstances indeed', the mind is 'something very mysterious'.[9]

Turing's Answer to the Mathematical Objection

In Chapter 11 Turing says (p. 451): 'The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect.' This remark might appear to cut to the heart of the matter. However, Turing expresses dissatisfaction with it, saying that the Mathematical Objection cannot 'be dismissed so lightly'. He goes on to broach a further line of attack on the argument, pointing out that humans 'often give wrong answers to questions', and it is this line of attack that he pursues in 'Intelligent Machinery, A Heretical Theory'.

In the quotation from Chapter 10 given above, Turing notes that the Mathematical Objection rests on a proviso that the machine is not allowed to make mistakes, and as he goes on to point out, 'the condition that the machine must not make mistakes . . . is not a requirement for intelligence' (p. 411). In 'Intelligent Machinery, A Heretical Theory' he suggests that the 'danger of the mathematician making mistakes is an unavoidable corollary of his power of sometimes hitting upon an entirely new method' (p. 472). Turing envisages machines also able to hit upon new methods: 'My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely. They will make mistakes at times, and at times they may make new and very interesting statements.'

Turing makes a similar point in Chapter 9 (pp. 393–4):

> [I]f a mathematician is confronted with such a problem [e.g. determining the truth or falsity of statements of the form '*p* is provable in such-and-such system'—Ed.] he would search around and find new methods of proof, so that he ought eventually to be able to reach a decision about any given formula. . . . I would say that fair play must be given to the machine. Instead of it sometimes giving no answer we could arrange that it gives occasional wrong answers. But the human mathematician would likewise make blunders when trying out new techniques. It is easy for us to regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy.

The use of heuristic search carries with it the risk of the computer producing a proportion of incorrect answers (see 'Artificial Intelligence'). This fact would have been very familiar to Turing from his experience with the bombe. Probably Turing was thinking of heuristic search when he wrote this, the earliest surviving

[9] 'Beyond the Doubting of a Shadow', section 13.2.

statement of his views concerning machine intelligence, in 'Proposed Electronic Calculator': 'There are indications however that it is possible to make the machine display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess.'[10]

In 'Intelligent Machinery, A Heretical Theory' Turing passes immediately from his remarks on the Mathematical Objection to a discussion of machine learning. This juxtaposition perhaps indicates that Turing's view was this: it is the possibility of a machine's learning *new* methods and techniques that ultimately defeats the Mathematical Objection. In the simplest possible case, the machine's tutor—a human mathematician—can just present the machine with a better method whenever the machine produces an incorrect answer to a problem. This new input in effect alters the machine's standard description, transforming it into a different Turing machine (see 'Computable Numbers: A Guide'). Alternatively a machine may itself be able to search around (albeit fallibly) for better methods. The search might involve the use of a random element. As in the preceding case, the standard description of the machine alters in consequence of the learning process, as the machine overwrites its previous algorithm with a successor. (As Turing says in Chapter 9: 'What we want is a machine that can learn from experience. The possibility of letting the machine alter its own instructions provides the mechanism for this.') Thus the learning machine may traverse the space of what in one of his letters to Newman (Chapter 4, p. 215) Turing calls 'proof finding' machines. In the same letter Turing says:

> One imagines different machines allowing different sets of proofs, and by choosing a suitable machine one can approximate 'truth' by 'provability' better than with a less suitable machine, and can in a sense approximate it as well as you please.

The learning machine successively mutates from one proof-finding Turing machine into another, becoming capable of wider sets of proofs as new, more powerful methods of proof are acquired.

The Future

Turing ends 'Intelligent Machinery, A Heretical Theory' with a vision of the future, now hackneyed, in which intelligent computers 'outstrip our feeble powers' and 'take control'. There is more of the same in Chapter 13. No doubt this is comic-strip stuff. Nevertheless, these images of Turing's reveal his profound grasp of the potential of the universal Turing machine at a time when the

[10] 'Proposed Electronic Calculator', National Physical Laboratory, 1945, 16 (National Physical Laboratory library; a digital facsimile of the original typescript is in The Turing Archive for the History of Computing <www.AlanTuring.net/proposed_electronic_calculator> (page reference is to the original typescript)).

only computers in existence were minuscule, and none but the most straightforward of tasks had been successfully programmed.[11]

Further reading

Benacerraf, P., 'God, the Devil, and Gödel', *Monist*, 51 (1967), 9–32.

Copeland, B. J., 'Turing's O-machines, Penrose, Searle, and the Brain', *Analysis*, 58 (1998), 128–38.

Gandy, R., 'Human versus Mechanical Intelligence', in P. Millican and A. Clark (eds.), *Machines and Thought: The Legacy of Alan Turing* (Oxford: Clarendon Press, 1996).

Lucas, J. R., 'Minds, Machines and Gödel', *Philosophy*, 36 (1961), 112–27.

——'Minds, Machines and Gödel: A Retrospect', in P. Millican and A. Clark (eds.), *Machines and Thought: The Legacy of Alan Turing* (Oxford: Clarendon Press, 1996).

Penrose, R., *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford: Oxford University Press, 1994).

Piccinini, G., 'Alan Turing and the Mathematical Objection', *Minds and Machines*, 13 (2003), 23–48.

Provenance

The text that follows is from a typescript entitled 'Intelligent Machinery, A Heretical Theory' and marked 'Typist's Typescript'.[12]

[11] In this chapter Turing speaks of the 'mechanic who has constructed the machine'. This is perhaps a glimpse of Turing's attitude toward Kilburn, Williams, and the other engineers who built the Manchester computer. Kilburn himself was hardly less dismissive of the logicians' contributions (for example in an interview with Christopher Evans in 1976, 'The Pioneers of Computing: An Oral History of Computing' (London: Science Museum)).

[12] The typescript is among the Turing Papers in the Modern Archive Centre, King's College, Cambridge (catalogue reference B 4). Turing's mother Sara included the text of 'Intelligent Machinery, A Heretical Theory' in her biography *Alan M. Turing* but unfortunately incorporated some errors (S. Turing, *Alan M. Turing* (Cambridge: Heffer, 1959), 128–34.) The present edition first appeared in B. J. Copeland (ed.), 'A Lecture and Two Radio Broadcasts on Machine Intelligence by Alan Turing', in K. Furukawa, D. Michie, and S. Muggleton (eds.), *Machine Intelligence 15* (Oxford: Oxford University Press, 1999).

Intelligent Machinery, A Heretical Theory

'You cannot make a machine to think for you.' This is a commonplace that is usually accepted without question. It will be the purpose of this paper to question it.

Most machinery developed for commercial purposes is intended to carry out some very specific job, and to carry it out with certainty and considerable speed. Very often it does the same series of operations over and over again without any variety. This fact about the actual machinery available is a powerful argument to many in favour of the slogan quoted above. To a mathematical logician this argument is not available, for it has been shown that there are machines theoretically possible which will do something very close to thinking. They will, for instance, test the validity of a formal proof in the system of Principia Mathematica, or even tell of a formula of that system whether it is provable or disprovable. In the case that the formula is neither provable nor disprovable such a machine certainly does not behave in a very satisfactory manner, for it continues to work indefinitely without producing any result at all, but this cannot be regarded as very different from the reaction of the mathematicians, who have for instance worked for hundreds of years on the question as to whether Fermat's last theorem is true or not. For the case of machines of this kind a more subtle argument is necessary. By Gödel's famous theorem, or some similar argument, one can show that however the machine is constructed there are bound to be cases where the machine fails to give an answer, but a mathematician would be able to. On the other hand, the machine has certain advantages over the mathematician. Whatever it does can be relied upon, assuming no mechanical 'breakdown', whereas the mathematician makes a certain proportion of mistakes. I believe that this danger of the mathematician making mistakes is an unavoidable corollary of his power of sometimes hitting upon an entirely new method. This seems to be confirmed by the well known fact that the most reliable people will not usually hit upon really new methods.

My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely. They will make mistakes at times, and at times they may make new and very interesting statements, and on the whole the output of them will be worth attention to the same sort of extent as the output of a human mind. The content of this statement lies in the greater frequency expected for the true statements, and it cannot, I think, be given an exact statement. It would not, for instance, be sufficient to say simply that the machine will make any true statement sooner or later, for an example of such a machine would be one which makes all possible statements sooner or later. We

know how to construct these, and as they would (probably) produce true and false statements about equally frequently, their verdicts would be quite worthless. It would be the actual reaction of the machine to circumstances that would prove my contention, if indeed it can be proved at all.

Let us go rather more carefully into the nature of this 'proof'. It is clearly possible to produce a machine which would give a very good account of itself for any range of tests, if the machine were made sufficiently elaborate. However, this again would hardly be considered an adequate proof. Such a machine would give itself away by making the same sort of mistake over and over again, and being quite unable to correct itself, or to be corrected by argument from outside. If the machine were able in some way to 'learn by experience' it would be much more impressive. If this were the case there seems to be no real reason why one should not start from a comparatively simple machine, and, by subjecting it to a suitable range of 'experience' transform it into one which was more elaborate, and was able to deal with a far greater range of contingencies. This process could probably be hastened by a suitable selection of the experiences to which it was subjected. This might be called 'education'. But here we have to be careful. It would be quite easy to arrange the experiences in such a way that they automatically caused the structure of the machine to build up into a previously intended form, and this would obviously be a gross form of cheating, almost on a par with having a man inside the machine. Here again the criterion as to what would be considered reasonable in the way of 'education' cannot be put into mathematical terms, but I suggest that the following would be adequate in practice. Let us suppose that it is intended that the machine shall understand English, and that owing to its having no hands or feet, and not needing to eat, nor desiring to smoke, it will occupy its time mostly in playing games such as Chess and GO, and possibly Bridge. The machine is provided with a typewriter keyboard on which any remarks to it are typed, and it also types out any remarks that it wishes to make. I suggest that the education of the machine should be entrusted to some highly competent schoolmaster who is interested in the project but who is forbidden any detailed knowledge of the inner workings of the machine. The mechanic who has constructed the machine, however, is permitted to keep the machine in running order, and if he suspects that the machine has been operating incorrectly may put it back to one of its previous positions and ask the schoolmaster to repeat his lessons from that point on, but he may not take any part in the teaching. Since this procedure would only serve to test the bona fides of the mechanic, I need hardly say that it would not be adopted in the experimental stages. As I see it, this education process would in practice be an essential to the production of a reasonably intelligent machine within a reasonably short space of time. The human analogy alone suggests this.

I may now give some indication of the way in which such a machine might be expected to function. The machine would incorporate a memory. This does not

need very much explanation. It would simply be a list of all the statements that had been made to it or by it, and all the moves it had made and the cards it had played in its games. This would be listed in chronological order. Besides this straightforward memory there would be a number of 'indexes of experiences'. To explain this idea I will suggest the form which one such index might possibly take. It might be an alphabetical index of the words that had been used giving the 'times' at which they had been used, so that they could be looked up in the memory. Another such index might contain patterns of men on parts of a GO board that had occurred. At comparatively late stages of education the memory might be extended to include important parts of the configuration of the machine at each moment, or in other words it would begin to remember what its thoughts had been. This would give rise to fruitful new forms of indexing. New forms of index might be introduced on account of special features observed in the indexes already used. The indexes would be used in this sort of way. Whenever a choice has to be made as to what to do next, features of the present situation are looked up in the indexes available, and the previous choice in the similar situations, and the outcome, good or bad, is discovered. The new choice is made accordingly. This raises a number of problems. If some of the indications are favourable and some are unfavourable what is one to do? The answer to this will probably differ from machine to machine and will also vary with its degree of education. At first probably some quite crude rule will suffice, e.g. to do whichever has the greatest number of votes in its favour. At a very late stage of education the whole question of procedure in such cases will probably have been investigated by the machine itself, by means of some kind of index, and this may result in some highly sophisticated, and, one hopes, highly satisfactory, form of rule. It seems probable however that the comparatively crude forms of rule will themselves be reasonably satisfactory, so that progress can on the whole be made in spite of the crudeness of the choice [of] rules.[1] This seems to be verified by the fact that engineering problems are sometimes solved by the crudest rule of thumb procedure which only deals with the most superficial aspects of the problem, e.g. whether a function increases or decreases with one of its variables. Another problem raised by this picture of the way behaviour is determined is the idea of 'favourable outcome'. Without some such idea, corresponding to the 'pleasure principle' of the psychologists, it is very difficult to see how to proceed. Certainly it would be most natural to introduce some such thing into the machine. I suggest that there should be two keys which can be manipulated by the schoolmaster, and which represent the ideas of pleasure and pain. At later stages in education the machine would recognise certain other conditions as desirable owing to their having been constantly associated in the past with pleasure, and likewise certain others as undesirable. Certain expressions of

[1] Editor's note. Words enclosed in square brackets do not appear in the typescript.

anger on the part of the schoolmaster might, for instance, be recognised as so ominous that they could never be overlooked, so that the schoolmaster would find that it became unnecessary to 'apply the cane' any more.

To make further suggestions along these lines would perhaps be unfruitful at this stage, as they are likely to consist of nothing more than an analysis of actual methods of education applied to human children. There is, however, one feature that I would like to suggest should be incorporated in the machines, and that is a 'random element'. Each machine should be supplied with a tape bearing a random series of figures, e.g. 0 and 1 in equal quantities, and this series of figures should be used in the choices made by the machine. This would result in the behaviour of the machine not being by any means completely determined by the experiences to which it was subjected, and would have some valuable uses when one was experimenting with it. By faking the choices made one would be able to control the development of the machine to some extent. One might, for instance, insist on the choice made being a particular one at, say, 10 particular places, and this would mean that about one machine in 1024 or more would develop to as high a degree as the one which had been faked. This cannot very well be given an accurate statement because of the subjective nature of the idea of 'degree of development' to say nothing of the fact that the machine that had been faked might have been also fortunate in its unfaked choices.

Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. To do so would of course meet with great opposition, unless we have advanced greatly in religious toleration from the days of Galileo. There would be great opposition from the intellectuals who were afraid of being put out of a job. It is probable though that the intellectuals would be mistaken about this. There would be plenty to do, [trying to understand what the machines were trying to say,][2] i.e. in trying to keep one's intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's 'Erewhon'.

[2] Editor's note. The words 'trying to understand what the machines were trying to say,' are handwritten and are marked in the margin 'Inserted from Turing's Typescript'.