Published in J. Copeland, J. Bowen, M. Sprevak & R. Wilson (Eds.) *The Turing Guide: Life, Work, Legacy* (2017), Oxford: Oxford University Press

mark.sprevak@ed.ac.uk

# Turing's model of the mind

Mark Sprevak
*University of Edinburgh*

13 November 2017

This chapter examines Alan Turing's contribution to the field that offers our best understanding of the mind: cognitive science. The idea that the human mind is (in some sense) a computer is central to cognitive science. Turing played a key role in developing this idea. The precise course of Turing's influence on cognitive science is complex and shows how seemingly abstract work in mathematical logic can spark a revolution in psychology.

Alan Turing contributed to a revolutionary idea: that mental activity is computation. Turing's work helped lay the foundation for what is now known as cognitive science. Today, computation is an essential element for explaining how the mind works. In this chapter, I return to Turing's early attempts to understanding the mind using computation and examine the role that Turing played in the early days of cognitive science.

## 1 Engineering versus psychology

Turing is famous as a founding figure in artificial intelligence (AI) but his contribution to cognitive science is less well known. The aim of AI is to create an intelligent machine. Turing was one of the first people to carry out research in AI, working on machine intelligence as early as 1941 and, as Chapters 29 and 30 explain, he was responsible for, or anticipated, many of the ideas that were later to shape AI.

Unlike AI, cognitive science does not aim to create an intelligent machine. It aims instead to understand the mechanisms that are peculiar to human intelligence. On the face of it, human intelligence is miraculous. How do we reason, understand language, remember past events, come up with a joke? It is hard to know how even to begin to explain these phenomena. Yet, like a magic trick that looks like a miracle to the audience, but which is explained by revealing the pulleys and levers behind the stage, so human intelligence could be explained if we knew the mechanisms that lie behind its production.

A first step in this direction is to examine a piece of machinery that is usually hidden from view: the human brain. A challenge is the astonishing complexity of the human brain: it is one of the most complex objects in the universe, containing 100 billion neurons and a web of around 100 trillion connections. Trying to uncover the mechanisms of human intelligence by looking at the brain is impossible unless one has an idea of what to look for. Which properties of the brain are relevant to intelligence? One of the guiding and most fruitful assumptions in cognitive science is that the relevant property of the brain for producing intelligence is the computation that the brain performs.

Cognitive science and AI are related: both concern human intelligence and both use computation. It is important to see, however, that their two projects are distinct. AI aims to create an intelligent machine that may or may not use the same mechanisms for intelligence as humans. Cognitive science aims to uncover the mechanisms peculiar to human intelligence. These two projects could, in principle, be pursued independently.

Consider that if one were to create an artificial hovering machine it is not also necessary to solve the problem of how birds and insects hover. Today, more than 100 years after the first helicopter flight, how birds and insects hover is still not understood. Similarly, if one were to create an intelligent machine, one need not also know how humans produce intelligent behaviour. One might be sanguine about AI but pessimistic about cognitive science. One might think that engineering an intelligent machine is possible, but that the mechanisms of human intelligence are too messy and complex to understand. Alternatively, one might think that human intelligence can be explained, but that the engineering challenge of building an intelligent machine is outside our reach.

In Turing's day, optimism reigned for AI and the cognitive-science project took a back seat. Fortunes have now reversed. Few AI researchers aim to create the kind of general, human-like, intelligence that Turing envisioned. In contrast, cognitive science is regarded as a highly promising research project.

Cognitive science and AI divide roughly along the lines of psychology versus engineering. Cognitive science aims to understand human intelligence; AI aims to engineer an intelligent machine. Turing's contribution to the AI project is well known. What did Turing contribute to the cognitive-science project? Did he intend his computational models as psychological models as well as engineering blueprints?

## 2 Building brainy computers

Turing rarely discussed psychology directly in his work. There is good evidence, however, that he saw computational models as shedding some light on human psychology.

Turing was fascinated by the idea of building a *brain-like* computer. His B-machines were inspired by his attempt to reproduce the action of the brain, as described in Chapter 29. Turing talked about his desire to build a machine to 'imitate a brain', to 'mimic the behaviour of the human computer', 'to take a man ... and to try to replace ... parts of him by machinery ... [with] some sort of "electronic brain" ', he claimed that 'it is not altogether unreasonable to describe digital computers as brains', and that 'our main problem [is] how to programme a machine to imitate a brain'.[1]

Evidently, Turing thought that the tasks in AI engineering and psychology were somehow related. What did he think was the nature of their relationship? We should distinguish three different things that he might have meant:

1. *Psychology sets standards for engineering success*. Human behaviour is where our grasp on intelligence starts. Intelligent behaviour is, in the first instance, known to us as something that humans do. One thing that psychology provides is a specification of human behaviour. This description can then be used in the service of AI by providing a benchmark for the behaviour of intelligent machines. Whether a machine counts as intelligent depends on how well it meets an appropriately idealized version of standards set by psychology. Psychology is relevant to AI because it specifies what is meant by intelligent behaviour. This connection seems peculiar to intelligent behaviour. One could, for instance, understand perfectly well what hovering is without knowledge of birds or insects.

2. *Psychology as a source of inspiration for engineering*. We know that the human brain produces intelligent behaviour. One way to tackle the AI engineering problem is to examine the human brain and take inspiration from it. Note, however, that the 'being inspired by' relation is a relatively weak one. Someone may be inspired by a design without understanding much about how that design works. Someone impressed by how birds hover may add wings to an

[1] Turing (2004b), p. 484; Turing (2004c), p. 445; Turing (2004d), p. 420; Turing (2004b), p. 482; Turing (2004d), p. 472.

artificial hovering machine. But even if this were successful, it would not mean the engineer knows how a bird's wings enable it to hover. Indeed, the way in which wings allows a bird to hover may not be the same as the way in which wings allow the engineer's artificial machine to hover – flapping may be an essential part in one case but not the other. An AI engineer might take inspiration from brains without knowing how brains work

3. *Psychology should explain human intelligence in terms of the brain's computational mechanisms*. Unlike the two previous claims, this involves the idea that the mechanisms of human thought are computational. The first two claims are compatible with this idea but they do not entail it. Indeed, the first two claims are silent about what psychology should, or should not, do. They describe a one-sided interaction between psychology and engineering with the influence going all from psychology to engineering: psychology sets the standards of engineering success or psychology inspires engineering. This claim is different: it recommends that psychology should adopt the computational framework of the AI engineering project. The way in which we explain human intelligence, and not just attempts to simulate it artificially, should be computational.

Did Turing make the third (cognitive-science) claim? Turing certainly gets close to it and, as we shall see in the final section, his work has been used by others in the service of that claim.

In the quotations above, Turing describes one possible strategy for AI: imitating the brain's mechanisms in an electronic computer. In order for such this strategy to work, one has to know which are the relevant properties of the brain to imitate. Turing says that the important features are not that 'the brain has the consistency of cold porridge' or any specific electrical property of nerves.[2] Rather, among the relevant features he cites the brain's ability 'to transmit information from place to place, and also to store it':[3]

> brains very nearly fall into [the class of electronic computers], and there seems to be every reason to believe that they could have been made to fall genuinely into it without any change in their essential properties.

On the face of it, this still has the flavour of a one-way interaction between AI engineering and psychology: which features of the brain *are relevant to* AI engineering? But unlike the claims above, this one-way interaction presupposes a specific view about how the human brain works: that the brain produces intelligent behaviour via (perhaps among other things) its computational properties. This is very close to

[2] Turing (2004a), p. 495; Turing (2004d), p. 420.
[3] Turing (2004d), p. 420.

the cognitive-science claim. Turing appears to be committed to something like to the third claim above (the cognitive-science claim) via his engineering strategy.

However, there is a problem with this reading of Turing. The key terms that Turing uses – 'reproduce', 'imitate', 'mimic', 'simulate' – have a special meaning in his work that is incompatible with the reading above. Those terms can be read as either 'strong' or 'weak'. On a strong reading, 'reproducing', 'imitating', 'mimicking', or 'simulating' means *copying that system's inner workings* – copying the equivalent of the levers and pulleys by which the system achieves its behaviour. On a weak reading, 'reproducing', 'imitating', 'mimicking', or 'simulating' means *copying the system's overall input-output behaviour* – reproducing the behaviour of the system, but not necessarily the system's method for doing so. The strong reading requires that an 'imitation' of a brain work in the same way as a real brain. The weak reading requires only that an imitation of a brain produce the same overall behaviour.

We assumed the strong reading above. In Turing's work, however, he tended to use the weak reading. Use of the weak reading is important to prove the computational results for which Turing is most famous (see Chapter 7). If the weak reading is the correct one, then the interpretation of Turing's words above is not correct. Imitating a brain does not require knowing how brains work – only knowing the overall behaviour brains produce. This falls squarely under the first relationship between psychology and engineering: psychology sets standards for engineering success. Imitating a brain – in the (weak) sense of reproducing the brain's overall behaviour – requires only that psychology specify the overall behaviour that AI should aim to reproduce. It does not require that psychology also adopt a computational theory about human psychology.

Is there evidence that Turing favoured the strong over the weak reading? Turing wrote to the psychologist W. Ross Ashby that:[4]

> In working on the ACE I am more interested in the possibility of producing models of the action of the brain than in practical applications to computing. … Thus, although the brain may in fact operate by changing its neuron circuits by the growth of axons and dendrites, we could nevertheless make a model, within the ACE, in which this possibility was allowed for, but in which the actual construction of the ACE did not alter, but only the remembered data, describing the mode of behaviour

[4] Letter from Turing to W. Ross Ashby, no date (Woodger papers (catalogue reference M11/99); a digital facsimile is in the Turing Archive for the History of Computing [www.alanturing.net/turing_ashby].

> applicable at any time.[5]

This appears to show that Turing endorsed something like the cognitive-science claim: he believed that the computational properties of the brain are the relevant ones to capture in a simulation of a brain. Unfortunately, it is also dogged by the same problem we saw previously. 'Producing a computational model of the action of the brain' can be given either a strong or a weak reading. It could mean *producing a model that works in the same way as the brain* (strong), or *producing a model that produces the same overall behaviour* (weak). Both kinds of computational model interested Turing and Ashby. Only the former would tell in favour of the cognitive-science claim.

Tantalisingly, Turing finished his 1951 BBC radio broadcast with:[6]

> The whole thinking process is still rather mysterious to us, but I believe that the attempt to make a thinking machine will help us greatly in finding out how we think ourselves.

The difficulty is that 'helping', like 'being inspired by', is not specific enough to pin the cognitive- science claim to Turing. There are many ways that the attempt to make a thinking machine might help psychology: the machines created might do useful number crunching, building the machines may teach us high-level principles that apply to all intelligent systems, building the machines may motivate psychology to give a specification of human competences. None of these would commit Turing to the cognitive-science claim.

Turing's writings are consistent with the cognitive-science claim but they do not offer unambiguous support for it. In the next section, we will see a clearer, but different, type of influence that Turing has had on modern-day cognitive science. In the final section, we will see how his computational models have been taken up and used by others as psychological models.

## 3 From mathematics to psychology

Turing proposed several computational models that have influenced psychology. Here I focus on only one: the Turing machine. Ostensibly, the purpose of the Turing machine was to settle questions about mathematics – in particular, the question of which mathematical statements can and cannot be proven by mechanical means.

[5] Letter from Turing to W. Ross Ashby, no date (Woodger papers (catalogue reference M11/99); a digital facsimile is in the Turing Archive for the History of Computing [www.alanturing.net/turing_ashby].

[6] Turing (2004b), p. 486.

We will see that Turing's model is good for another purpose: it can be used as a model of human thought. This spin-off benefit has been extremely influential.

A Turing machine is an abstract mathematical model of a human clerk. Imagine that a human being works by himself, mechanically, without undue intelligence or insight, to solve a mathematical problem. Turing asks us to compare this 'to a machine that is only capable of a finite number of conditions'.[7] That machine, a Turing machine, has a finite number of internal states in its head and an unlimited length of blank tape divided into squares on which it can write and erase symbols. At any moment, the machine can read a symbol from its tape, write a symbol, erase a symbol, move to neighbouring square, or change its internal state. Its behaviour is fixed by a finite set of instructions (a transition table) that specifies what it should do next in every circumstance (read, write, erase symbol, change state, move head).

Turing wanted to know which mathematical tasks could and could not be performed by a human clerk. Could a human clerk, given enough time and paper, calculate any number? Could a clerk tell us which mathematical statements are provable and which are not? Turing's brilliance was to see that these seemingly impossible questions about human clerks can be answered if we reformulate them to be about Turing machines. If one could show that the problems that can be solved by Turing machines are the same as the problems that can be solved by a human clerk, then any result about which problems a Turing machine can solve would carry over to a result about which problems a human clerk can solve. Turing machines can be proxies for human clerks in our reasoning.

It is easy to prove that the problems that a Turing machine can solve can also be solved by a human clerk. The clerk could simply step through the operations of the Turing machine by hand. Proving the converse claim – that the problems that a human clerk can solve could also be solved by a Turing machine – is harder. Turing offered a powerful informal argument for this second claim. Significantly, his argument depended on *psychological reasoning* about the human clerk:[8]

> The behaviour of the [clerk] at any moment is determined by the symbols which he is observing, and his 'state of mind' at that moment. We may suppose that there is a bound *B* to the number of symbols or squares that the [clerk] can observe at one moment. If he wishes to observe more, he must use successive observations. We will also suppose that the number of states of mind which need be taken into account is finite. The reasons for this are of the same character as those which restrict the number of symbols. If we admitted an infinity of states of

[7] Turing (2004e), p. 59.

[8] Turing (2004e), pp. 75–76.

> mind, some of them will be 'arbitrarily close' and will be confused.

Turing's strategy is to argue that the clerk cannot bring any more internal resources to bear in solving a problem than a Turing machine. Therefore, the class of problems that a clerk can solve is no larger than those of a Turing machine. In conjunction with the first claim above, this establishes the crucial claim that the problems that can be solved by Turing machines are exactly the same as those that can be solved by a human clerk.

Turing's argument is an exercise in weak modelling. His aim is to show that Turing machines and human clerks solve the same class of problems: they are capable of producing the same pattern of behaviour. His argument requires him to show that a Turing machine can copy the behaviour of the clerk and vice versa (weak modelling). It does not require him to show that the Turing machine reproduces that clerk's internal psychological mechanisms for generating his behaviour (strong modelling). Strong modelling goes beyond what was required by Turing's work on the *Entscheidungsproblem* but it is what we need for cognitive science.

One might conclude that there is nothing of further interest here for psychology. Yet, Turing's argument should give one pause for thought. Turing's argument requires that human clerks and Turing machines share at least *some* similarity in their inner working. They must have similar kinds of internal resources; otherwise, Turing's argument that the clerk's resources do not differ in kind from those of a Turing machine would not work. This suggests that a Turing machine is more than just a weak model of a human clerk. A Turing machine also provides a description, albeit rather high level and abstract, of the clerk's inner workings. In addition to capturing the clerk's outward behaviour, Turing machines also give some information about the levers and pulleys behind the clerk's behaviour.

## 4 Your brain's inner Turing machine

Does a Turing machine provide a psychologically realistic model of the mechanisms of the human mind? Turing never seriously pursued this question in print, but it has been taken up by others. The philosopher Hilary Putnam argued that a Turing machine is a good psychological model. Putnam claimed that a Turing machine is not only a good model of a clerk's mind while he is solving a mathematical task, it is a good model of other aspects of mental life.[9] According to Putnam, all human mental states (beliefs, desires, thoughts, imaginings, feelings, pains) should be understood as states of a Turing machine and its tape. All human mental processes (reasoning,

[9] Putnam (1975a); Putnam (1975c).

association, remembering) should be understood as computational steps of some Turing machine. Psychological explanation should be explanation in terms of the nature and operation of an inner Turing machine. Only when one sees the brain as implementing a Turing machine can one correctly see the contribution that the brain makes to our mental life. Putnam's proposal falls neatly under the cognitive-science claim identified above.

Putnam and others quickly became dissatisfied with the Turing machine as a psychological model.[10] It is not hard to see why. The human brain lacks any clear 'tape' or 'head', human mental states are not atomic states that change in a step-wise way over time, human psychology is not serial: it involves parallel mechanisms that cooperate or compete with each other. If the mind is a computer, it is unlikely to be a Turing machine.

The past fifty years have seen an explosion in the number and variety of computational models in psychology. State-of-the-art computational models of the mind look and work nothing like Turing machines. Among the most popular models are hierarchical recurrent connectionist networks that make probabilistic predictions and implement Bayesian inference.[11] The mechanisms of these computational models bear little resemblance to Turing machines. Yet, one might wonder, is there still something essentially right, albeit high level and abstract, about Turing machines as psychological models? And even if Turing machines do not model all aspects of our mental life, perhaps they provide a good model of some parts of our mental life.

Turing machines provide a good psychological model of at least one part our mental life: deliberate, serial, rule-governed inference – the capacity at work inside the head of the human clerk when he is solving his mathematical problems. In some situations humans deliberately arrange their mental processes to work in a rule-governed, serial way. They attempt to follow rules without using initiative, insight, or ingenuity, and without being disturbed by their other mental processes. In these situations, it seems that our psychological mechanisms approximate those of a Turing machine: our mental states appear step-wise, as atomic entities, and change in a serial fashion.

At a finer level of detail – and moving closer to the workings the brain – there is of course a more complex story to tell. Yet, as a 'high-level' computational model, the Turing machine is not bad as a piece of psychology. In certain situations, and at a high, abstract, level of description, our brains implement a Turing machine.

Modern computational models of the mind are massively parallel, exhibit complex and delicate dynamics, and operate with probability distributions rather than

[10] Putnam (1975b).

[11] Clark (2013).

discrete symbols. How can one square them with Turing machines? One way to integrate the two models is to use the idea that a Turing machine runs as a *virtual machine* on these models.[12] The idea is that a Turing machine arises, as an emergent phenomenon, out of some lower-level computational processes.[13] This idea should be familiar from electronic PCs: a high-level computation (in C# or Java) can arise out of lower-level computation (in assembler or microcode). High-level and low-level computational descriptions are both important when we explain how an electronic PC works. Similarly, we should expect that high-level and low-level descriptions will be important to explain how the human brain produces intelligence.

## 5 Conclusion

Turing has had a huge influence on cognitive science but, as we have seen, tracing the precise course of his influence is complex. In this chapter, we looked at two possible sources: Turing's discussion of how AI should be proceed, and the way in which Turing's computational models have influenced others. On the first score, we saw that Turing rarely talked about how AI should influence psychology, and that it is not easy to attribute to Turing the modern-day claim that human psychology should be computational. On the second, a clearer picture emerges. Turing's 1936 paper on the *Entscheidungsproblem* suggests that Turing machines are more than weak models of human psychology. Putnam and others took up this idea and proposed that Turing machines are strong models of human psychology. This idea remains influential today. Despite the wide range of exotic computational models in cognitive science, Turing machines still appear to capture a fundamental, albeit high-level truth about the workings of the human mind.

## Bibliography

Clark, A. (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and Brain Sciences* 36, pp. 181–253.

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown & Company.

Feldman, J. (2012). "Symbolic representation of probabilistic worlds". In: *Cognition* 123, pp. 61–83.

[12] See Dennett (1991).

[13] Zylberberg et al. (2011); Feldman (2012).

Putnam, H. (1975a). “Minds and machines”. In: *Mind, Language and Reality, Philosophical Papers, volume 2*. Cambridge: Cambridge University Press, pp. 362–387.

— (1975b). “Philosophy and our mental life”. In: *Mind, Language and Reality, Philosophical Papers, vol. 2*. Cambridge: Cambridge University Press, pp. 291–303.

— (1975c). “The mental life of some machines”. In: *Mind, Language and Reality, Philosophical Papers, volume 2*. Cambridge: Cambridge University Press, pp. 408–428.

Turing, A. M. (2004a). “Can automatic calculating machines be said to think?” In: *The Essential Turing*. Ed. by B. J. Copeland. Oxford: Oxford University Press, pp. 487–506.

— (2004b). “Can digital computers think?” In: *The Essential Turing*. Ed. by B. J. Copeland. Oxford: Oxford University Press, pp. 476–486.

— (2004c). “Computing machinery and intelligence”. In: *The Essential Turing*. Ed. by B. J. Copeland. Oxford: Oxford University Press, pp. 441–464.

— (2004d). “Intelligent machinery”. In: *The Essential Turing*. Ed. by B. J. Copeland. Oxford: Oxford University Press, pp. 395–432.

— (2004e). “On computable numbers, with an application to the *Entscheidungsproblem*”. In: *The Essential Turing*. Ed. by B. J. Copeland. Oxford: Oxford University Press, pp. 58–90.

Zylberberg, A. et al. (2011). “The human Turing machine: a neural framework for mental programs”. In: *Trends in Cognitive Sciences* 15, pp. 293–300.