mark.sprevak@ed.ac.uk

Turing's model of the mind

Mark Sprevak
University of Edinburgh

17 January 2014

Alan Turing contributed to a revolutionary idea: that mental activity is, at root, 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. The Human Brain Project, a giant European science project, recently awarded one billion euros, aims to understand the mind using this idea.[1] In this chapter, I return to Turing's early attempts to understanding the mind using computation. I examine the role Turing played in the early days of cognitive science.

1 Engineering versus psychology

Turing is famous as a founding figure in Artificial Intelligence, but his contribution to cognitive science is less well known. John McCarthy, who coined the term *Artificial Intelligence* (AI) in 1955, defined it as 'the science and engineering of making intelligent machines'. Turing was one of the first people to carry out research in this area, working on machine intelligence as early as 1941, and as Chapter 26 explained, he was responsible for, or anticipated, many of the ideas that were later to shape AI.

Cognitive science, unlike AI, does not aim to create an intelligent machine. Cognitive science aims instead to understand the mechanisms that are peculiar to human cognition. On the face of it, human intelligence seems miraculous. How do we reason, understand language, remember past events, come up with a joke? It is hard

1. See http://www.humanbrainproject.eu.

to know how even to begin to explain these phenomena. Yet, like a magic trick that seems miraculous to the audience, but which is explained by revealing the pulleys and levers behind the stage, so human intelligence could potentially 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 that then immediately confronts one is the human brain's astonishingly complexity. The human brain is one of the most complex objects in the universe. It contains a hundred billion neurons, and a mind-bogglingly complex web of close to a quadrillion connections. Trying to uncover the mechanisms of human intelligence by looking at the brain is impossible unless one has an idea what to look for. Which properties of the brain are relevant to producing 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 closely related: both concern human intelligence and both use computation. It is important to see, however, that the two fields are distinct. AI aims to create an intelligent machine that may, or may not, use the same mechanisms as humans. Cognitive science aims to uncover the mechanisms peculiar to human intelligence. The two projects may work in tandem, but they need not. Consider that if one aims to create a hovering machine, it is not necessary to also solve the problem of how birds and insects hover; today, more than 100 years after the first helicopter flight, it is still not fully understood how birds and insects hover. Similarly, if one aims to create an intelligent machine, one need not, at least in principle, explain how humans are intelligent. One might be sanguine about the former project, but pessimistic about the latter. For example, one might think that engineering an intelligent machine is doable, but that the mechanisms of human intelligence are too messy and complex for us ever to understand. Or, one might think that human intelligence, although messy and complex, can at least be investigated experimentally and is in broad outline explicable, but that building an intelligent machine depends on details that are too complex for engineers ever to master. In Turing's day, optimism reigned for AI, and the cognitive-science project took a backseat. Fortunes have now reversed. Few AI researchers aim at creating the kind of general intelligence that Turing envisioned. In contrast, cognitive science today is a highly promising research program and funded to the tune of billions of dollars as initiatives like the Human Brain Project show.

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

psychological theories as well as engineering strategy?

2 Building brainy computers

Turing rarely discussed psychology directly in his work. There is good evidence, however, that Turing saw computational models as shedding light on human psychology as well as a solution to the engineering problem of building an intelligent machine.

Turing was fascinated by the idea of building a brain-like computer. Turing's B-machines were inspired by his attempt to reproduce the action of the brain, as described in Chapter 30. 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"'. Turing claimed that 'it is not altogether unreasonable to describe digital computers as brains', that 'our main problem [is] how to programme a machine to imitate a brain'.[2]

Evidently, Turing thought the tasks of engineering and psychology were related. But what did Turing think was the nature of their relationship? We should distinguish three different things that Turing might have intended.

First, *psychology sets standards for engineering success*. Human behaviour is where our grasp on the notion of intelligence starts. Intelligent behaviour is, in the first instance, known to us as something that humans do. One thing that psychology provides us with is a specification of intelligent 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 idealised version of standards specified by psychology. Psychology is relevant to AI here because psychology specifies what is meant by intelligent behaviour. This connection seems peculiar to intelligent behaviour. One could, for instance, understand what *hovering* is perfectly well without knowledge of birds or insects.

Second, *psychology as a source of inspiration for engineering*. We know that the human brain produces intelligent behaviour. One way to tackle the engineering project is to examine the human brain. It is common for engineering to take inspiration from nature. Nevertheless, it should be emphasised that the *being-inspired-by* relation is relatively weak one. Someone may be inspired by a natural design without understanding, or having a worked-out view about, how that design works. For

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

example, someone impressed by how birds fly may add wings to an artificial flying machine. But even if this design proves fruitful, that does not mean the engineer knows how a bird's wings enable it to fly. The way in which a wing allows a bird to fly may not be the same as the way in which wing allows the engineer's artificial machine to fly—flapping, for example, may be essential part of the operation in one case but not the other. An engineer may take inspiration from brains without commitment to a computational view on how brains work.

Third, *psychology should explain human intelligence in terms of the brain's computational mechanisms*. Unlike the two previous claims, this claim does involve a commitment to the cognitive-science project. If this claim is right, the mechanisms of human thought are computational, and psychology should explain human intelligence in terms of computational mechanisms. The two claims above, although compatible with this claim, do not entail it. Indeed, the two claims above are silent about what psychology should, or should not, do. They only describe a one-sided interaction between psychology and engineering: psychology sets the standards of engineering success, or psychology inspires engineering. The cognitive-science claim is different. It recommends that psychology should be changed by the computational framework employed by the engineering project. The way in which we explain human cognition, as well as attempts to artificially simulate it, should be based on computational mechanisms.

Did Turing make the cognitive-science claim? Turing's work certainly has a great deal of affinity with this claim and, as we will see in the next section, his work has been used by others in the service of that claim. Turing himself appears to come close to asserting the cognitive-science claim at a number of points.

In his statements above, Turing describes an important strategy for AI: imitating the brain's mechanisms in an electronic computer. In order for such a strategy to work, one has to know the relevant properties of the brain for generating intelligence; otherwise, one would not know which aspects of the brain to reproduce. As Turing says, the relevant features are not that 'the brain has the consistency of cold porridge' or the fine-grained electrical properties of nerves.[3] The relevant feature, according to Turing, is the brain's ability 'to transmit information from place to place, and also to store it';[4] Turing says that '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.'[5] The essential properties of the brain for producing intelligence are, consequently, its computational properties. These are the properties responsible for human intelligent

3. Turing (2004a), p. 495; Turing (2004d), p. 420.
4. Turing (2004d), p. 420.
5. Turing (2004d), p. 412.

behaviour. On the face of it, this still has the flavour of a one-way interaction between engineering and psychology: which features of the brain *are relevant to engineering*? But unlike the two claims above, this one-way interaction presupposes a definite view on how the human brain works: the brain produces intelligent behaviour via its computational properties. This is precisely the cognitive-science claim. Turing appears, therefore, to be committed to the cognitive-science claim via his engineering strategy.

There is a problem, however. The key terms that Turing uses—'reproduce', 'imitate', 'mimic', 'simulate'—have a special meaning in Turing's work that pulls against the reading above. These terms can either have a *strong* or a *weak* reading. On the strong reading, *reproducing*, *imitating*, *mimicking*, or *simulating* a system means 'copying that system's inner workings'—copying the equivalent to the levers and pulleys by which the system achieves its behaviour. On the weak reading, *reproducing*, *imitating*, *mimicking*, or *simulating* means only 'copying the system's overall input–output behaviour'—reproducing the behaviour of the system, but not necessarily the system's particular tricks for doing so. The strong reading requires that an 'imitation' of a brain work in the same way as a brain; the weak reading requires only that an 'imitation' of a brain produce equivalent behaviour and be capable of solving the same tasks.

We assumed the strong reading above: for Turing to imitate, mimic, or simulate a human brain, he needed to make an assumption about how the human brain works. However, in Turing's work, he tended to use these terms primarily in their weak sense. Indeed, exclusive use of the weak sense is required in order to prove the computational results concerning the power of computers that Turing is most famous for, as we will see in the next section. If the weak sense of these terms is the correct one, then the interpretation above is not correct. Imitating a brain does not require knowing how brains work, only knowing which tasks brains solve. The latter 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 input–output behaviour—only requires psychology to specify the input–output behaviour that AI should aim to reproduce. It does not require that one accept the cognitive-science claim.

Is there evidence that Turing favoured the cognitive-science claim over the weak reading? Turing wrote to the psychologist W. Ross Ashby:

> 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 applicable at any time.[6]

This appears to show that Turing endorsed the cognitive-science claim: he believed that the computational properties of the brain are those that are essential to its operation in producing intelligent behaviour. Unfortunately, it is also dogged by the same problem we saw above. *Producing a computational model* can be given either a strong or weak reading. It could either mean producing a model that *works in the same way as the brain* (strong), or a model that merely *has the same overall behavioural profile* (weak). Both kinds of computational model would be of interest to Turing and Ashby. Only the former would tell in favour of the cognitive-science project.

Tantalisingly, Turing finished his 1951 BBC radio broadcast with:

> 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.[7]

The difficulty here is that the *helping* relation, like the *being-inspired-by* relation, is not specific enough to pin down the cognitive-science claim. There are innumerable ways in which the engineering project could help psychology: the machines created by the engineering project may facilitate psychological enquiry, the engineering project may teach us high-level principles that apply to all intelligent systems, the engineering project may motivate psychology to give a clear specification of human competences. None of these are the same as the cognitive-science claim.

Turing's writings are certainly consistent with the cognitive-science claim, and it may be natural to read that claim into his work. However, it is worth noting that his writings do not offer unambiguous support for this. In the next section, we will see a clearer form of influence Turing had on modern-day cognitive science. We will see how Turing's computational models were taken up by others and used as psychological models.

6. 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].

7. Turing (2004b), p. 486.

3 From mathematics to psychology

Turing proposed many computational models that have influenced psychology. Here I will focus only on the most famous of these models, the Turing machine. Turing devised the Turing machine as a mathematical abstraction of rule-governed behaviour. Ostensibly the purpose of the Turing machine was to settle questions in mathematical logic, in particular, the question of which mathematical statements can and cannot be proven by mechanical means. What we will see in this section is that Turing's model was good for another purpose: it could also be used in psychology as a model of human thought. This spin-off from Turing's mathematical work was extremely influential. While the Turing machine and its role in mathematical logic are described in Part 2 of this book, here we focus on the Turing machine's role in psychology.

A Turing machine is a mathematical abstraction of a human clerk. Suppose that such a human works 'mechanically', without 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'.[8] That machine, a Turing machine, has a finite number of internal states and an unlimited length of blank tape divided into squares on which it can write and erase symbols. At any moment, the Turing machine can read a symbol from its tape, write a symbol, erase a symbol, move to neighbouring square, or change its internal state. The Turing machine's behaviour is fixed by its finite set of instructions ('table') that specifies what it should do next (read, write, erase symbol, change state) when it reads a symbol and when it is in a particular state. Turing showed how to reason mathematically about the properties of Turing machines and how to prove results about the tasks that they can and cannot accomplish.

Turing wanted to know which tasks could be performed by a human clerk working mechanically. Could such a clerk, given enough time and paper, calculate any number? Could such a clerk, given enough time and paper, prove any true mathematical statement? Answering these questions in the abstract is hard; indeed, it is difficult to know where to begin. Turing's brilliance was to see that these seemingly intractable questions can be answered if we replace them with questions 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 the power of Turing machines would automatically carry over to human clerks. Turing machines are proxies for human clerks in our mathematical reasoning.

It is relatively straightforward to show that the problems that a Turing machine can solve can also be solved by a human clerk. The human clerk, given enough time

8. Turing (2004e), p. 59.

and paper, could simply step through the operation of the Turing machine by hand. The converse claim—problems that a human clerk can solve can also be solved by a Turing machine—is harder to establish. Turing offered a powerful informal argument for this claim. Significantly for our purposes, his argument depended on psychological reasoning about how the human clerk's mind works:

> 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 mind, some of them will be 'arbitrarily close' and will be confused.[9]

In this way, Turing argued that the clerk cannot bring any more internal resources to bear in solving a problem than a Turing machine. Hence, the class of problems that a clerk can solve can be no larger than those of a Turing machine. This is enough to establish his required claim that the problems that can be solved by Turing machines are the same as the problems that can be solved by a human clerk.

Turing's argument is a crucial exercise in what we called 'weak' modelling. Turing's aim was to show that Turing machines and human clerks solve the same class of problems. This only requires showing that a Turing machine mimics the input–output behaviour of a clerk ('weak modelling'), not that the Turing machine copies the clerk's peculiar internal mechanisms for doing so ('strong modelling'). The strong-modelling claim goes beyond what is required by Turing's 1936 paper, and Turing made no attempt to establish it. One might conclude from all this weak modelling that there is nothing of great interest here for psychology. Yet, the argument above should give one pause for thought. Turing's argument requires 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. That may lead one to wonder whether a Turing machine could be more than a weak model of a human. Perhaps a Turing machine also provides a description of the clerk's inner workings. In addition to capturing the clerk's outward behaviour, perhaps Turing machines also model the 'levers and pulleys' behind the clerk's behaviour. If correct, this would mean that Turing machines would not just be useful to mathematical logic, but also to psychology.

9. Turing (2004e), pp. 75–76.

4 Your brain's inner Turing machine

So does a Turing machine provide a psychologically-realistic model of the inner workings of a human clerk? Turing never pursued this question, but it has been taken up by others. Most notably, the philosopher Hilary Putnam argued that Turing machines are an accurate psychological model of the workings of the human mind. Putnam claimed that a Turing machine is not only a good model of the clerk's rule-governed problem-solving behaviour, a Turing machine is a good model of *all mental life*.[10] According to Putnam, all mental states (beliefs, desires, thoughts, imaginings, feelings, pains) should be understood as states of a Turing machine and its tape. Mental process involving these states (reasoning, association, remembering) should be understood as computational steps of that inner Turing machine. Psychological explanation should take the form of explanation in terms of the nature and operation of your inner Turing machine. Only once 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.

The context in which Putnam's model was proposed was quite different to that in psychology today. At the time, there was a lack of any kind of computational model of the human mind. The best account of the 'levers and pulleys' of our mental life was assumed to lie in the complex nitty-gritty of physiology—the particular details of physical brain processes—rather than in their computational properties. Putnam's target was a non-computational form of psychology. However, Putnam and others quickly became dissatisfied with the Turing machine as a model of human cognitive mechanisms.[11] It is not hard to see why Turing machines are not psychologically realistic. The human brain lacks any clear functional equivalent to a 'tape' or 'head', human mental states are not monolithic entities that change in a serial, step-wise way over time, and human psychology appears to involve different mechanisms that cooperate or compete with each other rather than a single mechanism that inexorably unfolds. If the mind is computational, it is unlikely that its computational architecture is that of a Turing machine.

The past fifty years have seen an explosion in the number, and sophistication, of computational models of the mind. Today, state-of-the-art computational models are a far cry from Turing machines. Among the most popular current models are hierarchical recurrent connectionist networks that make probabilistic predictions and implement Bayesian inference.[12] These probabilistic computational structures

10. Putnam (1975b, 1975d).
11. Putnam (1975c).
12. Clark (2013).

bear little resemblance in their inner working to Turing machines. Yet, one might still wonder if there is something essentially correct about Turing machines as a psychological model. Even if the Turing machine is not an accurate model of all aspects of our mental processing as Putnam had hoped, perhaps it nevertheless provides an accurate model of some limited portion of our mental life.

Current thinking is that Turing machines do provide a good psychological model of at least one aspect of our mental life: our conscious, serial, rule-governed thinking. This is precisely the mechanism at work in cases like that of the human clerk. In these cases, a human deliberately tries to arrange her mental processes to work in a rule-governed, serial way; she attempts to follow rules without using initiative, insight or ingenuity, and without being disturbed by her other mental processes. In these situations, it seems that human's psychological mechanisms do approximate those of a Turing machine: mental states appear as a single entity and change in a serial, step-wise fashion, and can be described as being governed by a single process rather than by competing mechanisms. At a finer level of detail—moving closer to the details of the brain—there are more complex and finer-grained computational stories to tell. Yet, as a high-level computational model, the Turing machine nevertheless still is useful to psychology. In certain contexts, and at least on some level, our brains implement a Turing machine. One way in which this result has been conveyed is that a Turing machine runs as a *virtual machine* on the human brain.[13]

Modern computational models of the mind, such as those described above, tend to be massively parallel, exhibit complex and delicate dynamics, and operate with probability distributions rather than discrete symbols. These computational models fare wonderfully at explaining unconscious mental processing. They also closely resemble the low-level computational details of the brain. In contrast, the Turing-machine model fares well at explaining conscious, serial, rule-governed, thinking, despite being quite distant from a brain-like architecture. Promising current research aims to bridge the gap between the two models and connect high-level Turing-machine description with the low-level computational architecture of the brain.[14] The general idea is that a Turing-machine arises naturally, as an emergent phenomenon, out of the action of low-level brain processes. An analogy could be drawn with an electronic PC: a high-level computational architecture (C# or Java) arises as an emergent phenomenon out of the joint action of a low-level computational architecture (assembler or microcode). Both low and high levels of computational description are important to explain different aspects of an electronic device's behaviour. It is therefore not surprising that psychology would continue to use both high-level Turing-machine-style description and low-level brain-like description in order to explain human behaviour.

13. See Dennett (1991).
14. Feldman (2012); Zylberberg et al. (2011).

In a rapidly-evolving field like cognitive science one might have guessed that by now Turing's computational models would have been discarded. Remarkably, this is not the case. Not only are Turing's general theoretical insights about computation intact, but, more extraordinary, Turing's specific computational models are regarded as fundamentally correct as a psychological description albeit in a limited domain. Turing machines provide a valuable high-level model of serial, rule-governed, thought processes. The Turing machine lives on in today's cognitive science as a virtual machine implemented on the human brain.

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 own discussion of how psychology should be conducted, and the way in which Turing's computational models have been used by others. On the first score, we saw that Turing rarely talked explicitly about how psychology should be conducted, and that it is not easy to attribute to Turing the modern-day cognitive-science claim based on his writings alone. On the second score—how Turing's computational writings have been used by others—a clearer picture of influence emerged. Turing's influential 1936 paper made it natural to ask whether the 'weak' computational modelling of humans that Turing established would lend itself to the 'strong' modelling of psychology. This idea, taken up by Putnam and others, remains influential today. Turing's legacy for cognitive science is immensely rich and complex; it is impossible to survey in its entirety here. An important part of that legacy, however, is that Turing machines capture a fundamental, and long-lasting, insight about the working of the human mind.

References

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

Copeland, B. J., ed. 2004. *The Essential Turing.* Oxford: Oxford University Press.

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

Feldman, J. 2012. 'Symbolic representation of probabilistic worlds'. *Cognition* 123:61–83.

Putnam, H. 1975a. *Mind, Language and Reality, Philosophical Papers, volume 2*. Cambridge: Cambridge University Press.

———. 1975b. 'Minds and machines'. In Putnam 1975a, 362–387.

———. 1975c. 'Philosophy and our mental life'. In *Mind, Language and Reality, Philosophical Papers, vol. 2,* 291–303. Cambridge: Cambridge University Press.

———. 1975d. 'The mental life of some machines'. In Putnam 1975a, 408–428.

Turing, A. M. 2004a. 'Can automatic calculating machines be said to think?' In Copeland 2004, 487–506.

———. 2004b. 'Can digital computers think?' In Copeland 2004, 476–486.

———. 2004c. 'Computing machinery and intelligence'. In Copeland 2004, 441–464.

———. 2004d. 'Intelligent machinery'. In Copeland 2004, 395–432.

———. 2004e. 'On computable numbers, with an application to the *Entscheidungsproblem*'. In Copeland 2004, 58–90.

Zylberberg, A., S. Dehaene, P. R. Roelfsema and M. Sigman. 2011. 'The human Turing machine: a neural framework for mental programs'. *Trends in Cognitive Sciences* 15:293–300.

28814c5 on 2014-01-17