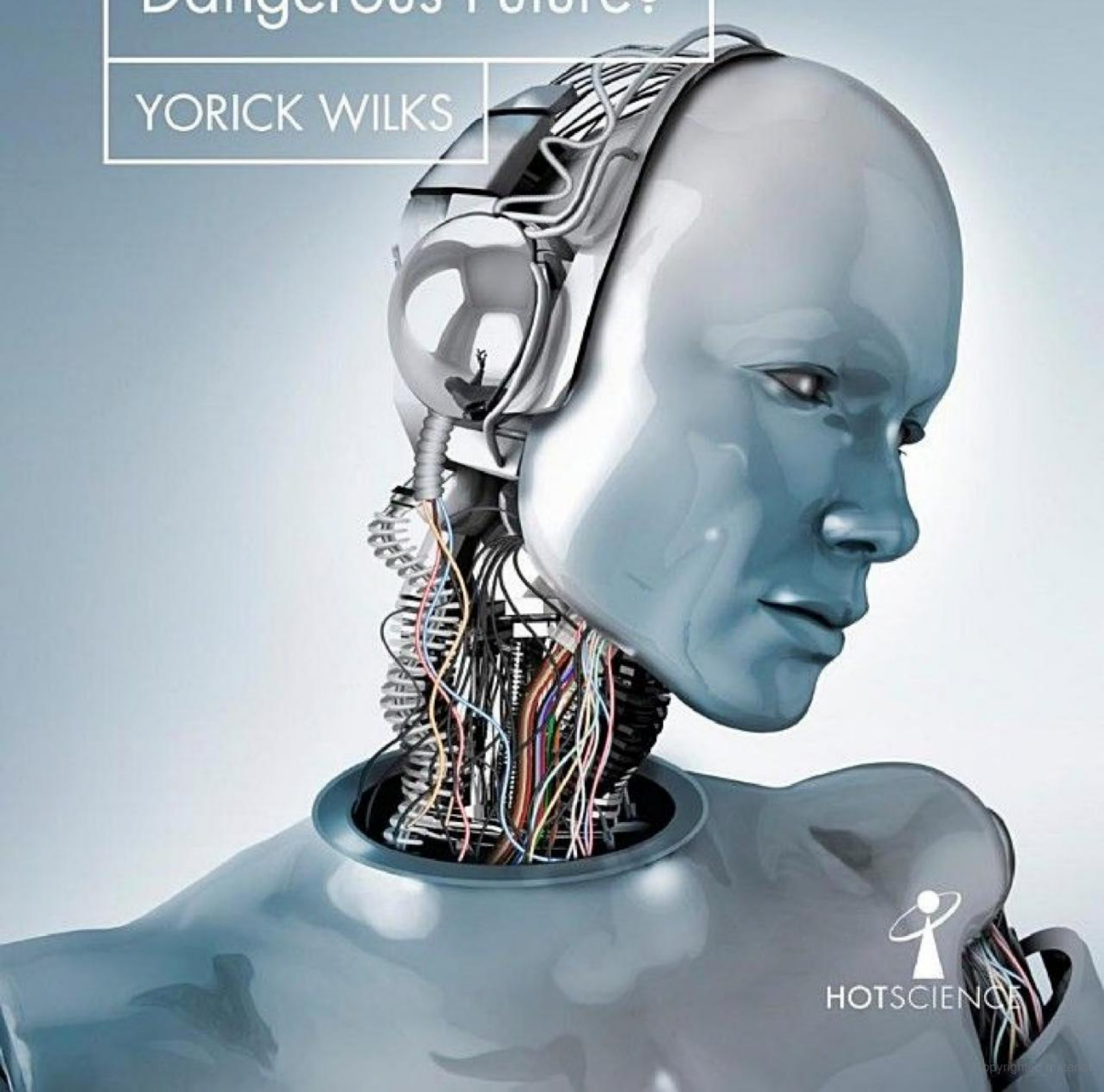
ARTIFICIAL INTELLIGENCE

Modern Magic or Dangerous Future?

YORICK WILKS



HOTSCIENCE

CONTENTS

ACKNOWLEDGEMENTS

I owe a debt of gratitude to Peet Morris who encouraged me to take on this book, as well as to Ken Ford for his continuing support over the years as Director of the Florida Institute of Human and Machine Cognition. I owe a great deal to the team at Icon (Duncan Heath, Brian Clegg and Rob Sharman) for their criticisms and suggestions. I also owe thanks to the many people who have made comments and criticisms on drafts of the book: Peet Morris, Patrick Hanks, Arthur Thomas, Robert Hoffman, Greg Grefenstette, Sergei Nirenburg, David Levy, John Tait, Tom Wachtel, Alexiei Dingli, Nick Ostler, Christine Madsen, Tomek Strzalkowski, Richard Weyhrauch, Derek Partridge, Angelo Dalli and Ken Lovesy. The errors, of course, are all my own.

1

SETTING OUT MY STALL: WHAT IS ARTIFICIAL INTELLIGENCE?

Wittgenstein wrote that a philosophy book could be written consisting entirely of jokes. In that spirit, an AI book could perhaps be written now consisting entirely of snippets from the daily news making claims about breakthroughs and discoveries. But something would be missing: how can we tell which of them are true, which really correspond to programs or are just science fiction fantasies? In this book, my aim is to describe the essence of AI now, but also to give an account of where it came from over a long period, and speculate about where it may be headed.

This very week one reads of a Belorussian woman who had programmed her dead fiancé's texts into a 'neural network' with which she could now talk posthumously. The same idea appeared in the *Black Mirror* episode 'Be Right Back' in 2013, and in an article I wrote in *Prospect* magazine in 2010 called 'Death and the Internet'. The same meme comes back all the time and is appealing, but did the Belorussian programmer really do anything serious? No one knows at the moment, but the need to distinguish research, pouring from companies and laboratories, from speculation, fantasy and fiction has never been greater, and I try to sort them out in this book.

The name 'artificial intelligence' was coined by American computer scientist John McCarthy, one of the handful of AI pioneers whose reputation still grows, for a 1956 workshop at Dartmouth College. But doubts about the phrase have grown since then and English code-breaker and computer scientist Donald Michie's previous version, 'Machine Intelligence', which it ousted, is making a comeback, and will be revived when the important journal *Nature Machine Intelligence* begins publication in 2019. That will be a badge of scientific respectability for a sometimes dubious field, where the word 'artificial' has come to have overtones of trickery. McCarthy said firmly that AI should be chiefly about getting computers to do things humans do easily and without thinking, such as seeing and talking, driving and manipulating objects, as well as planning our everyday lives. It should not, he said, be primarily about things that only a few people do well, such as playing chess or Go, or doing long division in their heads very fast, as calculators do. But

Michie thought chess was a key capacity of the human mind and that it should be at the core of AI. And the public triumphs of AI such as beating Kasparov – the then world champion – at chess, and more recently by playing world championship Go, have been taken as huge advances by those keen to show the inexorable advance of AI. But I shall take McCarthy's version as the working definition of AI for this book.

There can, then, be disputes about exactly what AI covers, as we shall see. I shall take a wide view in this book and try to give a quick and painless introduction to its history, achievements and aims – immediate and ultimate. The history is important, because although AI now seems everywhere, at least according to newspapers and media, and is pressing upon every human skill, it has actually been around for a long time and has lapped up around us very slowly. Here is a dramatic example: the road sign below was at the end of the driveway of the Stanford AI laboratory when I was there in the early 1970s.

This vehicle was rarely seen, and it consisted of four bicycle wheels with a wooden tray on top holding a radio aerial, a camera and a computer box. It could be steered by radio but sometimes ran itself round the driveway steered by the onboard computer. It was, though, far more significant than its absurd appearance. It was the beginning of the US government funded 'Moon Lander', later the 'Mars Lander', project, which was set up because it was known that vehicles on either body would have to be autonomous. That is to say, they would have to drive themselves, because they would be too far away to be radio controlled from Earth; they might fall down a crevasse in the time it took for a radio signal to reach them. That primitive vehicle ran almost 50 years ago, but is the father of all the autonomous vehicles now being tested on our roads and doing millions of miles a year.

Early setbacks to AI

It is important to see how long AI has been gestating, slowly but surely, even though it has been a bumpy ride with major setbacks. For example, in 1972 and 1973 AI suffered two major setbacks: the first was a book called *What Computers Can't Do*, by the philosopher Hubert Dreyfus. He called AI a kind of *alchemy* (forgetting for a moment that alchemy – an early form of chemistry which posited that metals could be transformed into each other – has actually turned out to be true in modern times with the discovery of nuclear transmutation!). Dreyfus's central point was that humans grew up, learning as they did so, and only creatures that did that could really understand as we do; that is to say, be true AI. Dreyfus's criticisms were rejected at the time by AI researchers, but actually had an effect on their work and understanding of what they were doing; he helped rejuvenate interest in machine learning as central to the AI project.

The following year, Sir James Lighthill, a distinguished control engineer, was asked by the British government to examine the prospects for AI. He produced a damning report the effect of which was to shut down research support in the UK for AI for many years, though some work was continued under other names such as 'Intelligent Knowledge Based Systems'. Lighthill's arguments about what counted as AI were almost all misconceived, as became clear years later. He himself had worked on automated landing systems for aircraft, a great technical success, and which we could easily now consider to be AI under the kind of definition given on page 2: the activity of simulating uniquely human activities and skills.

Lighthill considered that trying to model human psychology with computers was possible, but not AI's self-imposed task of just simulating human performances that required skill and knowledge. He was plainly wrong, of course – the existence of car-building robots, automated cars and effective machine translation on the web, as well as many AI achievements we now take for granted, all show that. Although a philosopher and an engineer respectively, Dreyfus and Lighthill had something in common: both saw that the AI project meant that computers had to have knowledge of the world to function. But for them, knowledge could not simply be poured into a machine as if from a hopper. AI researchers also recognised this need, yet believed such knowledge *could* be coded for a machine, though they disagreed about how. We shall revisit this topic – of knowledge and its representation – many times in the course of this book. Dreyfus thought you had to grow up and learn as we do to get such knowledge, but Lighthill intuited a form of something that AI researchers would describe as the 'frame problem' and he thought it insoluble.

The frame problem, put most simply, is that parts of the world around us 'update' themselves all the time depending on what kind of entity they are: if you turn a switch on, it stays on until you turn it off, but if it rains now, it very likely won't be raining in an hour's time. At some point it will stop. We all know this, but how is a computer to know that difference: that one kind of fact true now will stay true, but another will not be true some hours from now. We all learn as we grow up how the various bits of the world are, but can a computer know all that we know, so as to function as we do? At a key point in the film *Blade Runner*, a synthetic person, otherwise perfect, is exposed as such because it doesn't know that when a tortoise is turned over, it can't right itself.

The frame problem is serious and cannot be definitively solved, only dealt with by degrees. There have been many attempts, in AI and in computing generally, to prove that certain things *cannot be done*. Yet, in almost all cases these proofs turn out to be, not false, but useless because solutions can be engineered to get round the proofs and allow AI to proceed on its way. According to legend, Galileo, when before the Inquisition, where he was told firmly that the Earth could not possibly move, muttered under his breath the words *'Eppur si muove'* – 'and yet it moves'! Marvin Minsky at MIT, one of the great AI pioneers, once said that, yes, they ask for AI progress and it's hard to spot sometimes, but when you come back ten years later you are always astonished at how far it has moved.

The ghosts haunting AI over the years, telling its researchers what they cannot do, recall the 'proofs' given that machine translation (MT) was impossible. MT is another computer simulation of a very human skill that we could now consider a form of AI. In 1960, the Israeli philosopher Yehoshua Bar-Hillel argued that MT was impossible, because to translate language the system would have to have an enormous amount of world knowledge. His famous example was *The book was in the pen* where he argued that a computer would have to know a book could fit into a playpen but not into a writing pen, if it was to get the right sense of the word 'pen', and so translate the sentence out of English into some language where those were quite different words. This is an almost exact correspondence with the frame argument mounted against AI. Again, the everyday availability now of free MT of reasonable quality, from sources like Google Translate, shows how wrong Bar-Hillel was, though he was very influential at the time and widely believed.

I led a team in 1997 in New York that entered and won the Loebner competition, where an annual prize was awarded for the best computer conversationalist. The competition was set up so that journalists had short conversations at computer consoles, behind some of which were real people answering back, while behind others were the competitors, computer conversational programs. The job of the jury of journalists was to rate them all on 'humanness' and so decide on the best program of the year. How our team did this, and the kinds of tricks we used to fool the judges, included such things as making deliberate spelling mistakes to seem human, and making sure the computer responses came up slowly on the screen, as if being typed by a person, and not instantaneously as if read from stored data. Here was the start of one of the conversations in New York that actually convinced 'Judge 4' that our *Converse* program was a human:

> CONVERSE: Hi, my name is Catherine, what's yours?
>
> JUDGE 4: My name is Elliott.
>
> CONVERSE: Well, it's nice to meet you, Elliott. Did you see that story on CNN last night about the lesbian couple who came out at a White House party on Sunday?
>
> JUDGE 4: Yes, I did. I think it may all be a publicity stunt for Ellen [DeGeneres].

That output is now over twenty years old, and there hasn't been a great deal of advance since then in the performance of such 'chatbots'. This annual circus derived from Alan Turing's thoughts on intelligent machines in 1950, and his original test of how we might *know* a machine was thinking. His paper 'Computing Machinery and Intelligence' laid the groundwork for 70 years of discussion of the philosophical question 'Can a machine think?'.

Turing modelled his 'test' on a Victorian parlour game in which a contestant would ask questions, via folded notes passed from another room, with the aim of establishing whether the person answering the questions was a man or a woman. In the game Turing proposed, sex detection was still the aim, and if no one noticed when a computer was substituted, then the computer had in some sense *won*; it had been taken to be a person. The crucial point here is that the game was about men versus women – no one knew a computer might be playing. The irony, when we consider how his test has been adapted to events like the Loebner competition, is that Turing was not trying to say that computers did or ever would think: he was trying to shut down what he saw as useless philosophical discussion and present a practical test such that, if a machine passed it, we could just agree that they thought and so could stop arguing fruitlessly about the issue.

When we talk to others we never ask if they are machines or not. It doesn't make for a good conversation if you ask your friends that kind of thing. Nor did Turing think it would if we asked that of machines: that issue had to be *implicit*. Yet now in competitions such as the Loebner, the question 'Are you a computer?' has come out into the open and contestant machines are programmed to deal with it and give witty replies, as do the current commercial systems such as Alexa and Siri. But that is no longer any real test of anything except ingenuity.

My reason for mentioning the Loebner competition is that a curious feature of it is that the level of plausibility of the winning systems has not increased much over the last twenty years: systems that win don't usually enter again, as they have nothing left to prove. So new ones enter and win but do not seem any more fluent or convincing that those of a decade before. This is a corrective to the popular view that AI is always advancing all the time and at a great rate. As we shall see, some parts are, but some are quite static, and we shall need to ask why, and whether the answer lies in part in the optimistic promises researchers constantly make to the public and those who fund them.

Overpromising in AI – a persistent problem

It is important to come to grips with this issue because it is becoming harder to separate what AI has actually done from what it promises, and also from what the media think it promises. There are also science fiction worlds that are close to ours but hard to distinguish from reality. In the recent film *Her*, Scarlett Johansson's voice was given to a 'universal AI girlfriend' who seemed able to keep up close conversational relationships with millions of men worldwide. Since speaking and listening technologies such as Alexa are being sold all over the world, listen to their owners even when they are not attending to them, and then report their conversations back centrally, one can ask whether the public knows that Alexa exists but that the Johansson fiction does not? And the makers of sex robots are working hard to bring something like *Her* into existence. We shall need to be clear in what follows about what is known to work; what isn't – yet; and what may never work, no matter how hard we try.

Sorting these things out is made harder not only by company promises, made to sell products, but by researchers who have to constantly over-promise what they can do in order to win public research grants, a problem in the field since the Second World War. Already in the 1940s, when the capacity of the biggest computer was a millionth that of an iPhone, the papers were full of claims of 'giant brains', reasoning and thinking and just about to predict the weather for months to come. As early as 1946, the *Philadelphia Evening Bulletin* wrote of the technology at its local university that a '30-ton electronic brain at U of P thinks faster than Einstein'.

It was all nonsense of course, but there was real progress, too. Someone said recently that the most striking thing about today, to anyone who came here directly from the 1980s, would be that you could have something in your pocket that knew virtually everything there was to know. Think just how astonishing that is, let alone that it also makes phone calls. We marvel now at automatic cars, but computers have been landing planes without problems for nearly 40 years. One of my tasks here will be to convey what parts of AI are moving rapidly and which seem a little becalmed.

Two key questions

Two key questions will run, and hopefully be answered, throughout this book:

First, should AI be just using machines to imitate the performances humans give, or trying to do those things *the way we do them,* assuming we could know how our brains and bodies work? The two things can be quite different, with the first often thought of as engineering and the second as a way of doing psychology: explaining ourselves to ourselves by using computers. So, for example, some programs that determine the grammar structure of English sentences process them from right to left, i.e. backwards. They imitate our performance, but by methods we can be pretty certain differ from our own, and so could not be models of our own functioning.

It has long been a truism in AI thinking that, since the Wright brothers, aeroplanes fly but not with anything like the mechanism of flapping wings that birds use, and this example has been used to stress the difference between modelling the mechanism of evolution – of birds in this case – and really doing engineering. But more recently, the metaphor has reversed because it is now possible to build drones that *do* fly as birds do, and moreover to model in them the change of wing shape that enables many manoeuvres birds make but conventional planes cannot.

Secondly, should AI be based on building representations inside computers of how the world is, or should it just be manipulating numbers so as to imitate our behaviour? The current fashion in AI is for the second approach, called machine learning (ML), or even deep learning (DL), and many of the current news items in the media are about applications of this approach, such as the recent successes in diagnosing diseases or where a computer beat the best Go player in the world. Those are approaches based on numbers and statistics. But up until about 1990, the core AI approach used a form of logic to build representations – what I shall sometimes call ‘classical’ AI: structures representing things such as the layout of scenes or rooms. This is still how applications such as satnavs work, by internally examining structures of city streets to find the best route to drive. Such systems are not making statistical guesses about how the streets of London are connected.

This is an ongoing argument in AI research. When John McCarthy was offered statistical explanations back in the 1970s he would say, ‘But where do all these numbers come from?’ At the time, no one could tell him, but now they can, as we shall see in later chapters. One way of looking at this issue of those who want to represent things logically, versus those who think statistics a better guide to doing AI, is to remember how AI emerged; it was once bound up with a subject called cybernetics, a word now rarely used in the English-speaking world, though it is still used in Russia and in parts of Western Europe. Cybernetics was about reaching the goals of AI not with digital computers but with what were then called analogue computers, based not on logics but on continuous electrical processes, such as levels of current. Cybernetics produced things such as ‘smart’ home thermostats, and mechanical tortoises that could learn to plug themselves into wall sockets: they did not have representations in them at all. With the rise of classical, logic-based AI in the 1960s, in which reasoning was a central idea, cybernetics faded away as a separate subject. But the history of AI still has it jostling for space with other close disciplines such as control engineering (which pioneered planes that land automatically), pattern recognition (which introduced forms of machine vision) and statistical information retrieval.

There is nothing odd historically about subjects jostling against each other, disappearing in some cases (such as phrenology), or emerging from each other, as psychology and much of science did from philosophy in the 1800s. It’s a little like the prehistoric times when different tribes of humans – neanderthals, denisovans, *Homo sapiens* – co-existed, competed and interbred before one won out conclusively.

The case of information retrieval (IR) is important because of its link to the World Wide Web: the system of documents, images and video we now all have access to via our phones and computers. Google still dominates all search on the web, and the company’s founders Sergey Brin and Larry Page conceived their algorithm for searching it in the Stanford AI laboratory as part of PhDs they never finished. Yet, although coming from within an AI laboratory, Brin and Page’s search method was also directly within classic IR, but with a subtle twist I shall describe later on. The relevance of this to our big question is that IR, like cybernetics, does not deal in representations in a way that makes logic central, as ‘classical’ AI did.

Karen Spärck Jones was a Cambridge scientist who developed one of the basic tools for searching the web, and once argued that AI has much to learn from IR. Her main target was classical AI researchers, whom she saw as obsessed with content representations, when they should – according to her – have been making use of the statistical methods available in IR. Her arguments are very like those deployed by older cyberneticians, and more recently those who think machine learning is central to AI. Her questions to AI resolve to this crucial one: how can we capture the content of language except with its own words, or other words we use to explain them? Or, to put it another way, how could there be *other* representations of what language expresses that are not themselves language? This was a question that obsessed the philosopher Wittgenstein in the 1940s, and he seems to have believed language could not be represented by anything outside itself, or be compressed down into some logical coding. Here is a brief quotation from Spärck Jones in the 1990s that gives the flavour of her case that classical AI is simply wrong in thinking computers can reason with logical representations (what she calls the 'knowledge base') rather than by 'counting words' (another way of describing doing statistics with texts):

> The AI claim in its strongest form means that the knowledge base completely replaces the text base of the documents.

'Knowledge base' here means some logical structure a machine then uses to reason with, rather than the 'text base', that is, the original words themselves. This issue, of what it is that computers use as their basic representation of the world about which they reason, is still not settled. Most eye-catching developments in recent AI, from medicine, to playing Go, to machine translation on the Internet, are based on ideas closer to Spärck Jones and IR than to the logics and 'knowledge' on which AI was based for its first 50 years.

Will AI always be in digital computers or could it be in bodies?

A further question touched on towards the end of the book is about whether the basis of AI should be in digital computers at all, as it has been since cybernetics disappeared in the 1970s, or whether we shall reach AI not by copying how humans do things in computers but by merging computation with the biological, with real human or animal body tissues. For some, and this approach is more popular in Japan than in the West, this implies building up organic tissue-like structures that can perform, an approach one might parody as 'doing Frankenstein properly'. The alternative, more popular among American thinkers and entrepreneurs such as Elon Musk, is called ‘transhumanism’, the view that we could improve humans as they are now with artificial add-ons so that such beings gradually become a form of AI – and possibly immortal.

All these possibilities are full of religious and ethical overtones: of the creation myths of man in the Bible, of early artificial creatures such as the Golem of Prague, and of the ancient quest for immortality. I shall touch on serious questions such as this in Chapter 10.

The next chapters will describe the basic areas of artificial intelligence, including its relationship to the craft of computer programming, and we shall start with asking how important logic is to AI. McCarthy and others believed that AI was about making computer models of logical reasoning in machines and humans – an idea of the primacy of logic in thinking that goes back to the 1600s and Gottfried von Leibniz, the first man to say such things. I shall discuss the scope of machine logic and its decline and fall with the realisation that people do not seem to use logic much in everyday reasoning, nor even statistics.

2

HOW SHOULD ROBOTS THINK?: THE PLACE OF LOGIC IN AI

I used to give talks with titles such as ‘Keeping logic in its place’, which now sounds very presumptuous. I did that in reaction to the standard assumption in AI that the core of AI was *reasoning* done by computers, since that was also how human beings functioned. That assumption was almost universal throughout the last quarter of the twentieth century among the majority of AI researchers and developers: AI was then seen as the realisation of a dream going back to Aristotle, who thought humans were defined by their rationality, and most importantly to Leibniz, the seventeenth-century German philosopher.

Leibniz is most celebrated for inventing the calculus – albeit at the same moment as Newton – but he was also obsessed by the idea that if humans could only *calculate* together (the word he used for reasoning) then all problems, social and political, could be solved. He invented a logic of symbols – going beyond Aristotle, who worked with words to reason – and an artificial language in which to express thoughts and to reason about them. In a famous passage he wrote that if missionaries could only translate the Gospel into this language, heathens everywhere ‘could no more doubt its truths than the theorems of Euclid’. An ambitious project! Leibniz is a crucial step on the path to classical AI: the idea of a language of representation, not like the ordinary languages we speak, but one which could express what we want to say in some other, more exact way. As we saw in the last chapter, it was the assumption one could do that that Karen Spärck Jones claimed was the crucial fallacy of classical AI. Leibniz had no computers, of course, but he was aware of clockwork machines that could imitate humans; he thought reasoning was in some important sense logical *and mechanical.* In this spirit, later twentieth century logicians such as Bertrand Russell created much more complex logics of symbols that foreshadowed programming languages and the representations of AI. Russell also remarked that Leibniz was probably the cleverest man who had ever lived.

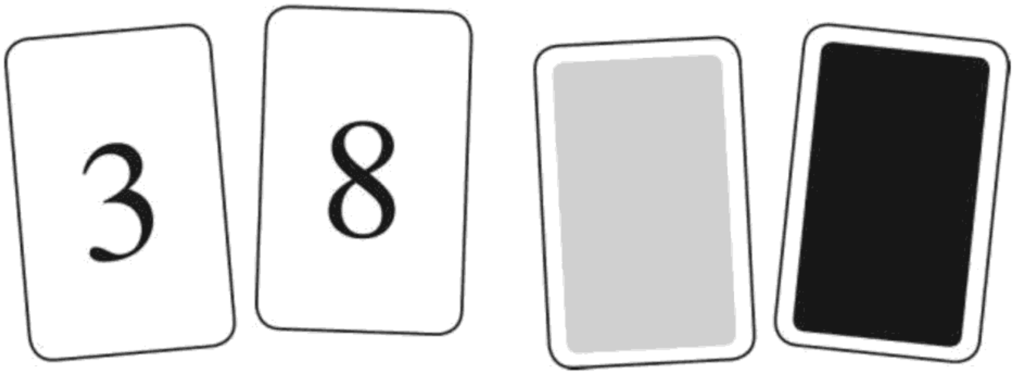
Pioneering AI researchers such as McCarthy also believed that even our simplest thought processes – not just chess and puzzles, but writing and planning our day – must rest on a system of logic running in our brains. Given that, the task of AI had to be to capture that reasoning in computers. There were three difficulties with this approach, usually known as 'theorem proving', since mechanical reasoning came down to knowing whether a particular sentence/theorem could be proved from another set of sentences or not.

Firstly, much psychology research has shown that people do not in fact reason logically in everyday life unless trained to do so through, for example, school exercises.

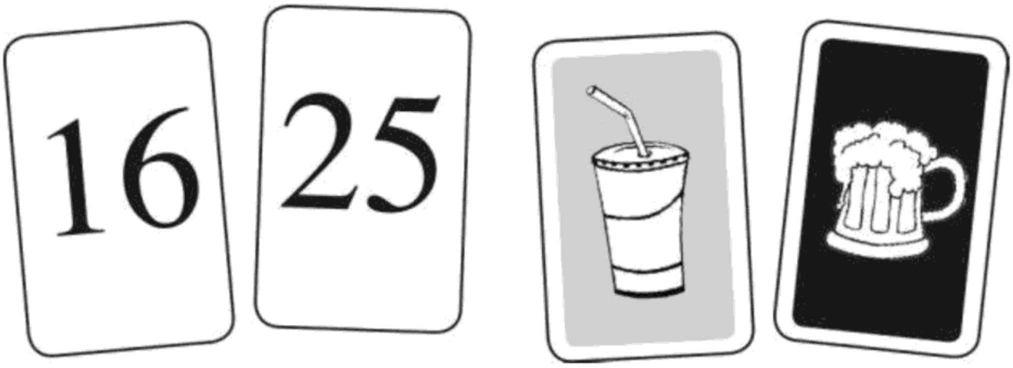
Secondly, research in logic often resulted in proofs that 'you can't do that', which some claimed to show that logic wasn't really up to the job of doing complete reasoning. The most famous example was Gödel's theorem showing that not all true sentences can be proved true, even though people can see they are true.

Thirdly, the practical results of theorem proving research over decades was not impressive in practical terms: it gave us few useful AI systems that worked.

As an illustration of the first point, we can look at some ingenious experiments of English psychologist Peter Wason in the 1960s, where subjects are shown cards, each of which has a colour on one side and a number on the other.



The question they are then asked is: Which cards must you turn over to test the claim that *if a card has an even number on one side it is grey on the other.* The answer is the 8 card and the black card, but only 10 per cent of people typically get that right. They fail to do the logical reasoning to show that, for example, if the grey card when turned over has an odd number on the back that doesn't show the claim wrong, because it was about cards with *even* numbers on one side. However, when the researchers put pictures of beer and soda on the cards, and ages such as 16 and 25 as the numbers, then subjects are much better at knowing which cards to turn to test the claim *if you are drinking alcohol then you must be at least 18.* In this case you must turn the 'beer' card and the 16.



The usual interpretation of this experiment is that people are not good at using logic in everyday problems without training, but if you change to a situation they know well, such as age and drinking, then they know what to do because they understand concrete situations, but not logic itself.

Another of Wason's demonstrations involved the sentence NO BRAIN INJURY IS TOO TRIVIAL TO BE IGNORED. He carried this about on a card and let people read it for themselves – he thought that if he read it aloud it would impose an interpretation on it. People were then asked to say whether it meant 'Treat all brain injuries' or 'Ignore all brain injuries'. Nearly all of them voted for the first, though in reality it means the second, as can be seen if you reverse its sense by substituting TREATED for IGNORED. You can then see that it does indeed mean the first, so it must have originally meant the second, strange though that seems, as the sentence is simple and contains no difficult or ambiguous words. What it does have, however, is a succession of powerful logical words such as NO, TOO and IG- (meaning *don't do it!*). Wason's explanation was that our mental processor finds this succession of logical moves too complicated to follow, so it simply gives up and plumps for the socially acceptable explanation, which is to treat injuries, not ignore them. Again, the message is that the human mind can't work very well with logic but goes directly for what it is used to or thinks acceptable.

Another recent blow to belief in the role of logic in human life has been the 'behavioural economics' pioneered by Daniel Kahneman, who showed conclusively that the standard economic model of humans as rational deciders of their own best interests was wildly false and that normal human behaviour was systematically irrational. A typical example concerned decisions on where to live in America: Kahneman showed that Californians and Midwesterners were equally happy on some standard scale, but *both* believed that Californians were happier, and prone to act on such a belief by moving house, against all the evidence.

Logic and what machines cannot do

I have already touched on the way AI has always been dogged by arguments about what computers cannot do, even in principle, and some of these concern logic and what can be proved. Gödel's theorem is the most famous of these and refers to two proofs Austrian logician Kurt Gödel gave in the 1930s about limits to the power of logic. The proofs were actually concerned with arithmetic but are taken to apply more generally. In the first theorem, Gödel showed that, for any logical system beyond the simplest, there would always be true statements it couldn't prove. This was a huge blow to those who thought logic was *complete*, meaning that all true things could be proved.

The second theorem showed that a logical system could not prove itself *consistent*, which is to say that it did not contain sentences that contradicted each other. The relevance of this to AI was thought to be that if logic and reasoning by computer were to be the core of AI, but you could never know for sure whether or not a given sentence could be proved true from other sentences, there would be a problem. It needs a bit of technical argument to get to that from the two theorems but imagine we have a set of sentences and want to prove one of them. For our purposes, let's refer to this sentence as p. Now imagine we added to the set a sentence that was the opposite of p – this can be termed NOTp. In logical terms, our original task of proving p from the original set is almost the same as proving that the second set (which contains both p and NOTp) is inconsistent – that is, it contains a contradiction. But if we cannot prove that there is that contradiction there – and Gödel's theorem says we cannot – then we cannot be sure we can prove the original sentence p. All that means is that you cannot be certain of proving everything that is true; it certainly doesn't mean that proving things with computers cannot work.

Again, we are in the situation where proving that things can't be done in all cases and everywhere doesn't mean that those things cannot be done as much as is necessary for practical purposes. What actually held up computer proofs was not this kind of demonstration of impossibility, but that the work of sorting through huge numbers of assumptions to find the ones relevant to what you were trying to prove was so time-consuming. But ingenious AI researchers found all kinds of shortcuts for doing just that.

Some philosophers jumped onto Gödel's proofs and tried to show they implied that computers could never have the capacities of humans – because there were things we could see were true but could not be proved by logic – and so full AI was impossible. But arguing that ignored the question of how humans knew those things were true in the first place. The argument assumed that computers must use logic to know things were true, but of course they could use other ways of finding out truth, just as we do, such as statistical guesses. We know most of what we know without proof and, since we do not know it by magic or occult powers, there must be such ways. Much of what we 'know' is what we were taught (and we believed), yet some of that will certainly be false, so we don't really know it. Why, then, should not machines do the same things we do to find out truths, such as accepting what teachers say?

It is tempting, faced with evidence such as Kahneman's on real human choices, to jump too far the other way and declare that logic is entirely the invention of schoolmen since Greek antiquity, and that it is all an educational illusion with no place in human life. But that would be quite wrong: modern civilisation rests on all kinds of rational processes and has done so since long before computers, right back to the construction of the pyramids and campaign plans of Roman armies. Some things simply have to be thought out and planned in fantastic detail: imagine building a jet engine from thousands of parts without a clear logical plan of what order to do things in.

There has always been another tradition in philosophy saying that the power of logic was overstated where everyday life was concerned. David Hume, arguably the greatest British philosopher, wrote in the eighteenth century: 'And if [ideas about facts] are apt, without extreme care, to fall into obscurity and confusion, the inferences are always much shorter in these disquisitions, and the intermediate steps much fewer than in the sciences.'

I take Hume to be saying that, outside the sciences, where deduction and logic have a real role, inferences are quite 'short', consisting of only a few steps. Moreover, they are informal steps rather than true logical steps, although Hume doesn't say they are informal because, two hundred years ago, he did not have the contrast we now have between human mental processes and formal logic or programming languages. But the kinds of inferences described above, with which people sort out which card to turn over to check a drinking age, suggest that he had in mind what we would now call informal inferences, which are brief and take place in something very like language itself.

A recent, much publicised success for AI illustrates this point. In 2011 IBM's WATSON system beat the best human contestants at the TV game *Jeopardy*, in which contestants are given an answer and have to guess the corresponding question. So, given the clue 'Its largest airport is named after a World War II hero and its second after a World War II battle', the correct response would be 'What are Chicago's airports named after?'. WATSON was not based on logic and reasoning except to a minor degree; it was a combination of IR – a skill we met already, here used to search through the enormous quantity of document data it held – and what is called natural language processing, or NLP. This is the area of AI which, like machine translation, deals with the manipulation and production of written language, a topic we shall discuss in much more detail in Chapter 5. WATSON would take the clue words such as 'airport' and 'World War II hero' and search in classic IR fashion for documents that contained them, and then try to isolate the sentences that might contain the answer. It would then make quite short inferences, in the style we associated with Hume above, to show that some string or words it had found did have the clue as its answer. WATSON was highly successful and a triumph not for logic at all but for IR and NLP.

WATSON received huge publicity and was touted by IBM as its entry point into medical AI: a way of providing doctors with expertise in response to questions put to a huge medical literature, a promise it has yet to fulfil. But there were two important misunderstandings about the WATSON experience. Firstly, it was not a breakthrough technology at all, but a refinement, over a long period, of state-of-the-art NLP technologies combined with IR: in particular, locating relevant sentence fragments and forming them into a proper response, then showing that that answered a question. This work had been done over decades as part of DARPA-funded projects. DARPA, originally ARPA, is the US Defense Advanced Research Projects Agency, the main historical supporter of AI research. It was also the US government agency that had funded the Internet, and much of US NLP.

Secondly, although WATSON performed well in the competition in real time, its victory was almost certainly due not only to its processing speed but to the speed with which it could press the buzzer first. There is a real delay in humans because of the time taken by the nerve signal from the brain to the finger on the buzzer; but WATSON had no brain and no finger and could answer in far fewer milliseconds than the humans. That was almost certainly decisive and, in hindsight, WATSON should have been handicapped for fairness.

WATSON was a qualified success and showed the importance of limited uses of logic, in this case combined with complex text processing. Another fact that made possible the later revival of statistical and non-representational machine learning methods in the 1990s was that more thoroughgoing logical methods than WATSON produced very few products that captured the imagination of either researchers or the public. However, a scaled-down version of the logic program was renamed 'expert systems' and did have some limited success, chiefly in the contexts of science and medicine, and also in specialised areas such as using logic to optimise the distribution of computers across a given floor space.

Like the early robot vehicle and the original Google algorithm, the most striking expert systems came from Stanford: in this case the laboratory of Ed Feigenbaum, and with names such as DENDRAL and MYCIN. The former was a system to identify new forms of organic compounds that might be possible and be worth synthesising in the laboratory. As the name *expert systems* suggests, these efforts were not the computer modelling of everyday life skills such as language and vision, but were based on coding into rules the knowledge of scientific and medical experts, so as ultimately to outperform them, which these systems did in the 1970s and 1980s. MYCIN was the first truly effective computer system for diagnosing infectious diseases, and the first AI system to encounter the inevitable and powerful resistance from the medical profession to the challenge it posed, which it has taken nearly 40 years to overcome. It is vital to stress that such expert systems were not pure, theoretical theorem-provers at all, but worked with shortcuts, usually called *heuristics* – procedures that trimmed and made practical the searches for evidence to prove conclusions. Furthermore, the power of such systems was in the expert knowledge they contained rather than the logic itself. As Feigenbaum put it: ‘intelligent systems derive their power from the knowledge they possess rather than from the specific formalisms and inference schemes they use.’ This work also gave impetus to a separate psychological discipline called ‘knowledge elicitation’, a method for getting experts to reveal their knowledge so it could be coded – sometimes knowledge they didn’t even know they had.

We should look briefly at another key aspect of logic that has been important in creating AI representations of the world and our knowledge of it. A fundamental principle of logic since Leibniz has been the ability to substitute alternative names for the same thing in a logical form *without changing the meaning or truth of the whole*, whether we call that a formula or an expression or a sentence. So, if Joe is the name of Jane’s father and I write ‘Joe is married to Ann’ that will be just as true if I substitute the other words for ‘Joe’ to get the sentence ‘Jane’s father is married to Ann’ because ‘Joe’ and ‘Jane’s father’ are the same person. A large part of logic is devoted to the problems that arise with the simple move of adding in human belief to create sentences such as ‘I believe Joe is married to Ann’. The problem arises because that may *not* be true at the same time as ‘I believe Jane’s father is married to Ann’ is true, although they are same person. It’s simply that I don’t happen to know the *name* of Jane’s father.

A whole industry has grown up within logic to deal with this issue, called *opacity* (meaning contexts you can’t see through). But what is important to AI, in its desire to model human representations of the world, is that it must have some practical way to deal with this issue of belief if it is to model individual humans conversing with each other. Such individuals will almost certainly have different knowledge states and different, inconsistent, beliefs about the world, such as who is married to whom. It is obvious we are able to talk to people who do not share all our beliefs – about democracy as much as about family members – and one way of describing that is to say we have models of each other in our minds and consult these so as to communicate. I talk to you based on a model I have of you, and of what you may or may not believe that differs from what I believe. A teacher has to have this model in order to teach a student who, by definition, does not yet know or believe the things that the teacher does. Similarly, a doctor will have to talk to a patient who refers to a pain in his stomach but points to the wrong place, because the patient has a wrong belief about where his stomach actually is (as many do, believing it to be lower than its real position).

In later chapters I shall discuss individuals and their own knowledge and belief in relation to the Internet, and will assume that AI can construct and manipulate representations, or *models*, of the beliefs of individual people in order to have computers talk to them effectively. The notion of a machine *having a belief* will therefore be an important one for AI, though not one I want to use carelessly. I do not want to say an ATM has a belief about how much is in my bank account: I want rather to say it just has data about me. But if a computer could have a model of me in which it could see that I thought I had more in my account that I actually do, and could therefore correct me gently about that, I might want to say it did have a belief *because it could contrast two points of view of my account and see how they differed.* This idea of machine belief as having alternative points of view is an important inheritance from logic – one that AI needs – but, interestingly, not one that logicians ever developed much for themselves. AI is not just a handmaid of logicians, but has sometimes developed logical ideas further within itself. Yet, as we shall see, logic has been vital to AI in another way: in the construction of special programming languages to construct AI systems. The languages PROLOG and LISP – which I shall discuss in Chapter 4 – have been crucial to the development of AI programs and were developed directly from ideas in logic. The AI–logic relation is a complicated and mixed one that can be characterised as being about practicality versus theory, but which is also one of great indebtedness to two-and-a-half thousand years of intense thought.

COPYRIGHT

Published in the UK and USA in 2019
by Icon Books Ltd, Omnibus Business Centre,
39–41 North Road, London N7 9DP
email: info@iconbooks.com
www.iconbooks.com

Sold in the UK, Europe and Asia
by Faber & Faber Ltd, Bloomsbury House,
74–77 Great Russell Street,
London WC1B 3DA or their agents

Distributed in the UK, Europe and Asia
by Grantham Book Services,
Trent Road, Grantham NG31 7XQ

Distributed in the USA
by Publishers Group West,
1700 Fourth Street, Berkeley, CA 94710

Distributed in Australia and New Zealand
by Allen & Unwin Pty Ltd,
PO Box 8500, 83 Alexander Street,
Crows Nest, NSW 2065

Distributed in South Africa
by Jonathan Ball, Office B4, The District,
41 Sir Lowry Road, Woodstock 7925

Distributed in India by Penguin Books India,
7th Floor, Infinity Tower – C, DLF Cyber City,
Gurgaon 122002, Haryana

Distributed in Canada by Publishers Group Canada,
76 Stafford Street, Unit 300
Toronto, Ontario M6J 2S1

ISBN: 978–178578–516–0
eISBN: 978–178578–517–7

Typeset in Iowan by Marie Doherty

Printed and bound in Great Britain
by Clays Ltd, Elcograf S.p.A.