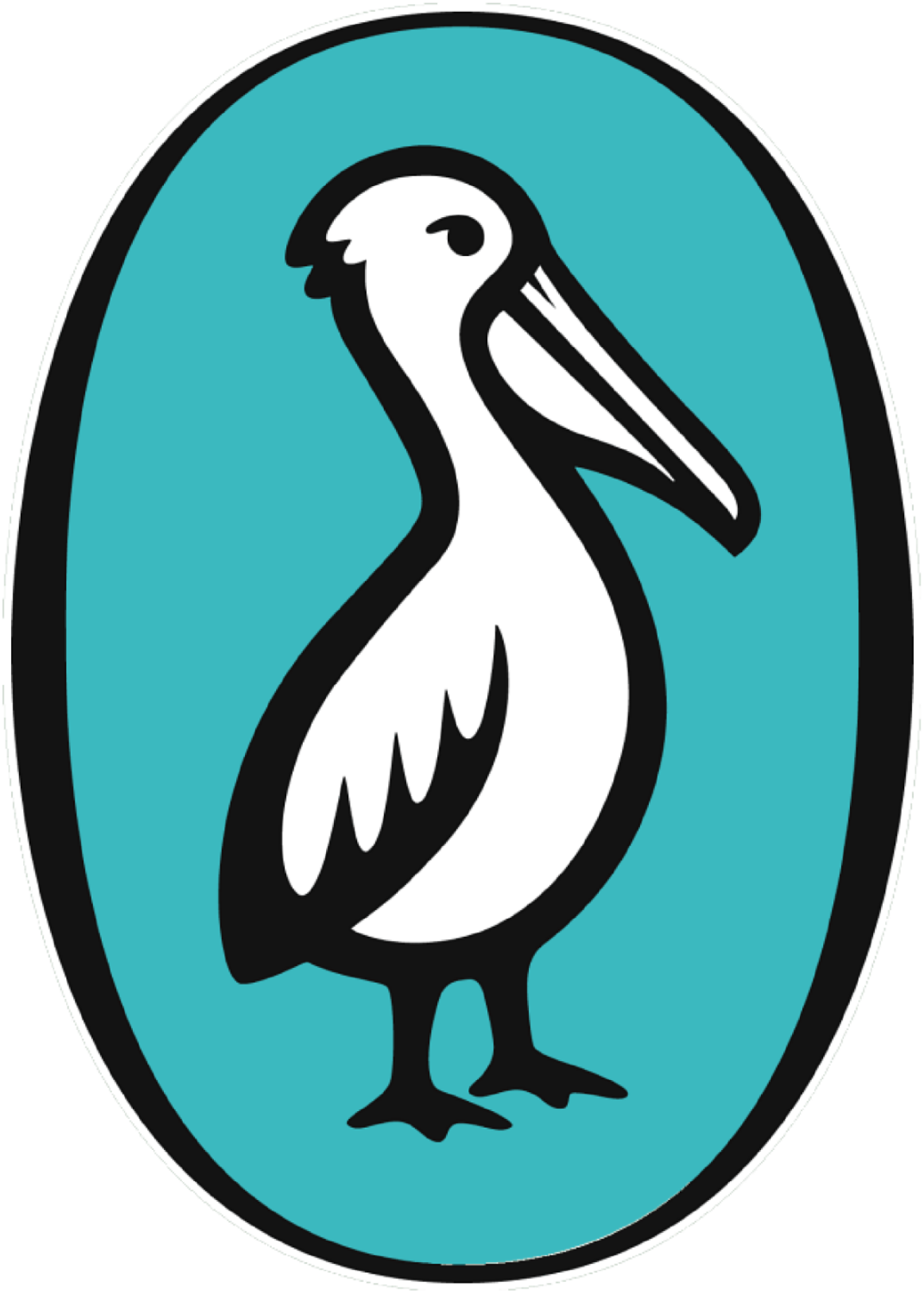

Contents

ACKNOWLEDGEMENTS

INTRODUCTION

PART ONE: WHAT IS AI?

CHAPTER 1
Turing's Electronic Brains

PART TWO: HOW DID WE GET HERE?

CHAPTER 2
The Golden Age

CHAPTER 3
Knowledge Is Power

CHAPTER 4
Robots and Rationality

CHAPTER 5
Deep Breakthroughs

PART THREE: WHERE ARE WE GOING?

CHAPTER 6
AI Today

CHAPTER 7
How We Imagine Things Might Go Wrong

CHAPTER 8
How Things Might Actually Go Wrong

CHAPTER 9
Conscious Machines?

GLOSSARY

APPENDIX A: UNDERSTANDING RULES

APPENDIX B: UNDERSTANDING PROLOG

APPENDIX C: UNDERSTANDING BAYES' THEOREM

APPENDIX D: UNDERSTANDING NEURAL NETS

FURTHER READING

NOTES

INDEX

About the Author

Michael Wooldridge is a professor of Computer Science and Head of the Department of Computer Science at the University of Oxford, where he is a Fellow of Hertford College. He has been an AI researcher since 1989, and has published more than 400 scientific articles on the subject. From 2014 to 2016, he was President of the European Association for AI, and from 2015 to 2017 he was President of the International Joint Conference on AI (IJCAI). He lives in Oxford with his wife and two children.

For the Principal, Fellows and Scholars of Hertford College in the University of Oxford

Acknowledgements

I am enormously grateful to my agent Felicity Bryan and my editor Laura Stickney for their support, encouragement and advice throughout this project. And, I should add, for their (ahem) good-natured patience.

Special thanks to: Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy for their permission to use the 'panda/gibbon' images that appear in Figure 17, from their paper 'Explaining and Harnessing Adversarial Examples'; Peter Millican for permission to adapt some material from a paper we wrote together; Subbarao Kambhampati for useful feedback on some of my main arguments, and pointing me towards Pascal's wager; Nigel Shadbolt for discussion, encouragement and fantastic anecdotes about the history of AI; and Reid G. Smith for his permission to adapt some of his teaching materials relating to the MYCIN system.

I am grateful to those colleagues, students and unlucky acquaintances who read drafts of the book at short notice and gave enormously valuable feedback. Thanks here to Ani Calinescu, Tim Clement-Jones, Carl Benedikt Frey, Paul Harrenstein, Andrew Hodges, Matthias Holweg, Will Hutton, Graham May, Aida Mehonic, Peter Millican, Steve New, André Nilsen, James Paulin, Emma Smith, Thomas Steeples, André Stern, John Thornhill and Kiri Walden. Naturally, the many errors which surely remain are entirely my responsibility.

I am honoured to acknowledge the support of the European Research Council under Advanced Grant 291528, which funded me and my research group from 2012 to 2018. My work was made infinitely more productive and immensely more pleasurable by an energetic, brilliant and endlessly supportive research group led by Julian Gutierrez and Paul Harrenstein. I very much hope we will continue to collaborate far into the future.

INTRODUCTION

About halfway through writing this book, I was having lunch with a colleague.

'What are you working on?' she asked me.

This is a standard question for academics – we ask it of each other all the time. I should have been ready for it, and had an impressive answer ready to hand.

'Something a bit different. I'm writing a popular science introduction to artificial intelligence.'

She snorted. 'Does the world really need *yet another* popular science introduction to AI? What's the main idea then? What's your new angle?'

I was crestfallen. I needed a clever comeback. So I made a joke.

'It's the story of AI through failed ideas.'

She looked at me, her smile now faded. 'It's going to be a bloody long book, then.'

Artificial intelligence (AI) is my life. I fell in love with AI as a student in the mid-1980s, and I remain passionate about it today. I love AI, not because I think it will make me rich (although that would be nice), nor because I believe it will transform our world (although, as we will see in this book, I believe it *will* do so, in many important ways). I love AI because it is the most endlessly fascinating subject I know of. It draws upon and contributes to an astonishing range of disciplines, including philosophy, psychology, cognitive science, neuroscience, logic, statistics, economics and robotics. And ultimately, of course, AI appeals to fundamental questions about the human condition, and our status as *homo sapiens* – what it means to be human, and whether humans are unique.

What AI Is and Isn't

My first main goal in this book is to tell you what AI is – and, perhaps more importantly, what it is not. You might find this a little surprising, because it may seem obvious to you what AI is all about. Surely, the long-term dream of AI is to build machines that have the full range of capabilities for intelligent action that people have – to build machines that are self-aware, conscious and autonomous in the same way that people like you and me are. You will probably have encountered this version of the AI dream in science-fiction movies, TV shows and books.

This version of AI may seem intuitive and obvious, but as we will see when we try to understand what it really means, we encounter many difficulties. The truth is we don't

remotely understand what it is we want to create, or the mechanisms that create it in people. Moreover, it is by no means the case that there is agreement that this really *is* the goal of AI. In fact, it is fiercely contentious – there isn't even any consensus that this kind of AI is feasible, let alone desirable.

For these reasons, this version of AI – *the grand dream* – is difficult to approach directly, and although it makes for great books, movies and video games, it isn't in the mainstream of AI research. Of course, the grand dream raises quite profound philosophical questions – and we will discuss many of these in this book. But beyond these, much of what is written about this version of AI is really nothing more than speculation. Some of it is of the lunatic fringe variety – AI has always attracted crackpots, charlatans and snake oil salesmen as well as brilliant scientists.

Nevertheless, the public debate on AI, and the seemingly never-ending press fascination with it, is largely fixated on the grand dream, and on alarmist dystopian scenarios that have become a weary trope when reporting on AI (AI will take all our jobs; AI will get smarter than us, and then it will be out of control; super-intelligent AI might go wrong and eliminate humanity). Much of what is published about AI in the popular press is ill-informed or irrelevant. Most of it is garbage, from a technical point of view, however entertaining it might be.

In this book, I want to change that narrative. I want to tell you about what AI *actually* is, what AI researchers *actually* work on and *how they go about it*. The reality of AI for the foreseeable future is very different to the grand dream. It is perhaps less immediately attention-grabbing, but it is, as I will show in this book, tremendously exciting in its own right. The mainstream of AI research today is focused around getting machines to do specific tasks which currently require human brains (and also, potentially, human bodies), and for which conventional computing techniques provide no solution. This century has witnessed important advances in this area, which is why AI is so fêted at present. Automated translation tools are one example of an AI technology that was firmly in the realm of science fiction 20 years ago, which has become a practical, everyday reality within the past decade. Such tools have many limitations, but they are successfully used by millions of people across the globe every day. Within the next decade, we will see high-quality real-time spoken-word language translation, and augmented reality tools that will change the way we perceive, understand and relate to the world we live in. Driverless cars are a realistic prospect, and AI looks set to have transformative applications in healthcare, from which we will all stand to benefit: AI systems have proven to be better than people at recognizing abnormalities such as tumours on X-rays and ultrasound scans, and wearable technology, coupled with AI, has the potential to monitor our health on a continual basis, giving us advance warnings of heart disease, stress and even dementia. *This* is the kind of thing that AI researchers actually work on. *This* is what excites me about AI. And this is what the AI narrative should be about.

To understand what AI today is, and why AI is for the most part not concerned with the grand dream, we also need to understand why AI is hard to create. Over the past 60 years, huge amounts of effort (and research funding) have flowed into AI, and yet, sadly, robot butlers are not likely any time soon. So, why has AI proved to be so difficult? To understand the answer to this question, we need to understand what computers are and what computers can do, at their most fundamental level. This takes us into the realm of some of the deepest questions in mathematics, and the work of one of the greatest minds of the twentieth century: Alan Turing.

The History of AI

My second main goal in this book is to tell you the story of AI from its inception. Every story must have a plot, and we are told there are really only seven basic plots for all the stories in existence, so which of these best fits the story of AI? Many of my colleagues would dearly like it to be 'Rags to Riches', and it has certainly turned out that way for a clever (or lucky) few. For reasons that will become clear later, we could also plausibly view the AI story as 'Slaying the Beast' – the beast, in this case, being an abstract mathematical theory called computational complexity, which came to explain why so many AI problems proved fearsomely hard to solve. 'The Quest' would also work, because the story of AI is indeed rather like that of medieval knights on a quest to find the Holy Grail: full of religious fervour, hopeless optimism, false leads, dead ends and bitter disappointments. But, in the end, the plot that best fits AI is 'Fall and Rise', because, only 20 years ago, AI was a rather niche area with a somewhat questionable academic reputation – but since then it has risen to be the most vibrant and celebrated area in contemporary science. It would be more accurate, though, to say that the plot to the AI story is 'Rise and Fall and Rise and Fall and Rise'. AI has been the subject of continuous research for more than half a century, but during this time AI researchers have repeatedly claimed to have made breakthroughs that bring the dream of intelligent machines within reach, only to have their claims exposed as hopelessly over-optimistic in every case. As a consequence, AI is notorious for boom-and-bust cycles – there have been at least three such cycles in the past four decades. At several points over the past 60 years, the bust has been so severe that it seemed like AI might never recover – and yet, in each case, it did. If you imagine that science is all about orderly progress from ignorance to enlightenment, then I'm afraid you are in for a bit of a shock.

Right now, we are in boom times yet again, and excitement is at fever pitch. At such a time of fevered expectations it is, I think, essential that the troubled story of AI is told and told again. AI researchers have, more than once, thought they had discovered the magic ingredient to propel AI to its destiny – and the breathless, wide-eyed discussion of current AI has precisely the same tone. Sundar Pichai, CEO of Google, was reported to have claimed that 'AI is one of the most important things humanity is working on. It is more profound than, I dunno, electricity or fire.'[1] This followed an earlier claim by Andrew Ng, to the effect that AI was 'the new electricity'.[2] It is important to remember what came of such hubris in the past. Yes, there have been real advances, and yes, they are cause for celebration and excitement – but they do not take us to the end of the road: to conscious machines. As an AI researcher with 30 years of experience, I have learned to be obsessively cautious about the claims made for my field: I have a very well-developed sense of scepticism about claims for breakthroughs. What AI needs now, perhaps more than anything, is a very large injection of humility. I am reminded of the story from Ancient Rome of the *auriga* – a slave who would accompany a triumphant general on his victory march through the city, repeatedly whispering in the general's ear the Latin phrase *memento homo* – 'Remember, you are only human'.

The second part of this book therefore tells the story of AI, warts and all, in broadly chronological order. The story of AI begins just after the creation of the first computers following the Second World War. I will take you through each of the boom periods – starting with the Golden Age of AI, a period of unbridled optimism, when it seemed for a while that rapid progress was being made on a broad range of fronts; next through the 'knowledge era', when the big idea was to give machines all the knowledge that we have about our world; and, more recently, the behavioural period, which insisted that robots should be centre stage in AI, taking us up to the present time. In each case, we'll meet the ideas and the people that shaped AI in their time.

The Future for AI

While I believe it is important to understand the present hubbub about all things AI within the context of a long history of failed ideas (and to temper one's excitement accordingly), I also believe there is real cause for optimism about AI right now. There *have* been genuine scientific breakthroughs, and these, coupled with the availability of 'big data' and very cheap computer power, have made possible in the past decade AI systems that the founders of the field would have hailed as miraculous. So, my third main goal is to tell you what AI systems can actually do right now, and what they might be able to do soon – and to point out their limitations.

This leads me into a discussion of fears about AI. As I mentioned above, the public debate about AI is dominated by dystopian scenarios such as AI taking over the world. Recent advances in AI *do* raise issues that we should all be concerned about, such as the nature of employment in the age of AI, and how AI technologies might affect human rights – but discussions about whether robots will take over the world (or whatever) take the headlines away from these very important, very real concerns. So, once again, I want to change the narrative. I want to take you through the main areas of concern, and to signpost as clearly as I can what you *should* be afraid of, and what you should not.

Finally, I want us to have some fun. So, in the final chapter, we will return to the grand dream of AI – conscious, self-aware, autonomous machines. We will dig into that dream in more detail, and ask what it would mean to realize it, what such machines would be like – and whether they would be like us.

How to Read This Book

The remainder of this book is structured around these overarching goals: to tell you what AI is and why it is hard; to tell you the story of AI – the ideas and people that drove it through each of its boom periods; and, finally, to showcase what AI can do now, and what it can't, and to talk a little about the long-term prospects for AI – the road to conscious machines.

One of the pleasures of writing this book has been to cast off the usual important but tiresome conventions of academic writing. Thus, there are relatively few references, although I give citations for the most important points.

Not only have I avoided giving extensive references, I've also steered clear of technical details – by which I mean *mathematical* details. My hope is that, after reading this book, you will have a broad understanding of the main ideas and concepts that have driven AI throughout its history. Most of these ideas and concepts are, in truth, highly mathematical in nature. But I am acutely aware of Stephen Hawking's dictum that every equation in a book will halve its readership. For those that feel up to the challenge, I have included some appendices that dig a little deeper into some of the technical ideas, and I also include some points for further reading.

The book is *highly* selective. I really had no choice here: AI is an enormous field, and it would be utterly impossible to do justice to all the different ideas, traditions and schools of thought that have influenced AI over the past 60 years. Instead, I have tried to single out what I see as being the main threads which make up the complex tapestry that is the story of AI.

Finally, I should caution that this is not a textbook. Reading this book will not equip you with the skills to start a new AI company, or to join the staff of Google or Facebook. What you will gain from this book is an understanding of what AI is and where it might be heading. I hope that, after reading it, you will be properly informed about the *reality* of AI – and that, after reading it, you will help me to change the narrative.

Oxford, May 2019

PART ONE

What is AI?

CHAPTER 1

Turing's Electronic Brains

I propose to consider the question, 'Can machines think?'
–– Alan Turing (1950)

Every story needs to start somewhere, and for AI we have many possible choices, because the dream of AI, in some form or other, is an ancient one.

We could choose to begin in Classical Greece, with the story of Hephaestus, blacksmith to the gods, who had the power to bring metal creatures to life.

We could begin in the city of Prague in the 1600s, where, according to legend, the head rabbi created the Golem – a magical being fashioned from clay, intended to protect the city's Jewish population from anti-Semitic attacks.

We could begin with James Watt in eighteenth-century Scotland, designing the 'Governor', an ingenious automatic control system for the steam engines he was building, thereby laying the foundations for modern control theory.

We could begin in the early nineteenth century with the young Mary Shelley, cooped up in a Swiss villa during a spell of bad weather, creating the story of Frankenstein to entertain her husband, the poet Percy Bysshe Shelley, and their family friend, the notorious Lord Byron.

We could begin in London in the 1830s with Ada Lovelace, estranged daughter of the same Lord Byron, striking up a friendship with Charles Babbage, curmudgeonly inventor of mechanical calculating machines, and inspiring the brilliant young Ada to speculate about whether machines might ultimately be creative.

We could equally well begin with the eighteenth-century fascination with automata – cunningly designed machines that gave some illusion of life.

We have many possible choices for the beginning of AI, but for me the beginning of the AI story coincides with the beginning of the story of computing itself, for which we have a pretty clear starting point: King's College, Cambridge, in 1935, and a brilliant but unconventional young student called Alan Turing.

Cambridge, 1935

apply them – they are nothing more than *recipes*, which can be followed by rote. All we need to do to find the answer is to follow the recipe *precisely*.

Since we have a technique that is guaranteed to answer the question (given sufficient time), we say that questions in the form 'Is *n* a prime number?' are **decidable**. I emphasize that all this means is that, whenever we are faced with a question in the form 'Is *n* a prime number?', we know that we can definitely find the answer if we are given sufficient time: we follow the relevant recipe, and eventually we will get the correct answer.

Now, the *Entscheidungsproblem* asks whether all mathematical decision problems like those we saw above are decidable, or whether there are problems for which there is *no* recipe for finding the answer – no matter how much time you are prepared to put in.

This is a very fundamental question – it asks whether mathematics can be reduced to merely following recipes. And answering this fundamental question was the daunting challenge that Turing set himself in 1935 – and which he triumphantly resolved, with dizzying speed.

When we think of deep mathematical problems, we imagine that any solution to them must involve complex equations and long proofs. And sometimes, this is indeed the case – when the British mathematician Andrew Wiles famously proved Fermat's Last Theorem in the early 1990s, it took years for the mathematical community to understand the hundreds of pages of his proof, and become confident that it was indeed correct. By these standards, Turing's solution to the *Entscheidungsproblem* was positively eccentric.

Apart from anything else, Turing's proof is short and comparatively accessible (once the basic framework has been established, the proof is really just a few lines long). But most importantly, to solve the *Entscheidungsproblem*, Turing realized that he needed to be able to make the idea of a recipe that can be followed precisely exact. To do this, he invented a mathematical problem-solving machine – nowadays, we call these **Turing machines** in his honour. A Turing machine is a mathematical description of a recipe, like the one for checking prime numbers mentioned above. All a Turing machine does is to follow the recipe it was designed for. I should emphasize that, although Turing called them 'machines', at this point they were nothing more than an abstract mathematical idea. The idea of solving a deep mathematical problem by *inventing a machine* was unconventional, to say the least – I suspect many mathematicians of the day were mystified.

Turing machines are very powerful beasts. Any kind of mathematical recipe that you might care to think of can be encoded as a Turing machine. And, if all mathematical decision problems can be solved by following a recipe, then for any decision problem, you should be able to design a Turing machine to solve it. To settle Hilbert's problem, all you had to do was show that there was some decision problem that could not be answered by any Turing machine. And that is what Turing did.

His next trick was to show that his machines could be turned into *general-purpose* problem-solving machines. He designed a Turing machine that will follow *any* recipe that you give it. We now call these general-purpose Turing machines **Universal Turing Machines**:[4] and a computer, when stripped down to its bare essentials, is simply a Universal Turing Machine made real. The programs that a computer runs are just recipes, like the one for prime numbers that we discussed above.

Although it isn't central to our story, it is worth at least mentioning how Turing settled the *Entscheidungsproblem* using his new invention – apart from the fact that it was extraordinarily ingenious, it also has some bearing on the question of whether AI is ultimately possible.

His idea was that Turing machines could be programmed to *answer questions about other Turing machines.* He considered the following decision problem: given a Turing

PART TWO

How did We Get Here?

CHAPTER 2

The Golden Age

Although Turing's article 'Computing Machinery and Intelligence', which introduced the Turing test, made what we now recognize as the first substantial scientific contribution to the discipline of AI, it was a rather isolated contribution, because AI as a discipline simply did not exist at the time. It did not have a name, there was no community of researchers working on it, and the only contributions at the time were speculative conceptual ones, such as the Turing test – there were no AI systems. Just a decade later, by the end of the 1950s all that had changed: a new discipline had been established, with a distinctive name; and researchers were able to proudly show off the first tentative systems demonstrating rudimentary components of intelligent behaviour.

The next two decades were the first boom in AI. There was a flush of optimism, growth and apparent progress, leading to the era called the **Golden Age of** AI, from about 1956 to 1974. There had been no disappointments yet; everything seemed possible. The AI systems built in this period are legends in the AI canon. Systems with quirky, geeky names like SHRDLU, STRIPS and SHAKEY – short names, all in upper-case, supposedly because those were the constraints of computer file names at the time (the tradition of naming AI systems in this way continues to the present day, although it has long since ceased to be necessary). The computers used to build these systems were, by modern standards, unimaginably limited, painfully slow and tremendously hard to use. The tools we take for granted when developing software today did not exist then and indeed could not have run on the computers of the time. Much of the 'hacker' culture of computer programming seems to have emerged at the time. AI researchers worked at night because then they could get access to the computers that were used for more important work during normal office hours; and they had to invent all sorts of ingenious programming tricks to get their complicated programs to run at all – many of these tricks subsequently became standard techniques, with their origins in the AI labs of the 1960s and 1970s now only dimly remembered, if at all.[1]

But by the mid-1970s, progress on AI stalled, having failed to progress far beyond the earliest simple experiments. The young discipline came close to being snuffed out by research funders and a scientific community that came to believe AI, which had promised so much, was actually going nowhere.

In this chapter, we'll look at these first two decades of AI. We'll look at some of the key systems built during this period, and discuss one of the most important techniques developed in AI at the time – a technique called a 'search', which to the present day

remains a central component of many AI systems. We'll also hear how an abstract mathematical theory, called computational complexity and developed in the late 1960s and early 1970s, began to explain why so many problems in AI were fundamentally hard. Computational complexity cast a long shadow over AI.

We'll begin with the traditional starting point of the Golden Age: the summer of 1956, when the field was given its name by a young American academic by the name of John McCarthy.

The First Summer of AI

McCarthy belonged to that generation of academics who created the modern technological USA. With almost casual brilliance, throughout the 1950s and 1960s, he invented a range of concepts in computing that are now taken so much for granted that it is hard to imagine that they actually had to be invented. One of his most famous developments was a programming language called **LISP**, which for decades was the programming language of choice for AI researchers. At the best of times computer programs are hard to read, but even by the frankly arcane standards of my profession, LISP is regarded as bizarre, because in LISP (all (programs (look (like this)))) – generations of programmers learned to joke that LISP stood for 'Lots of Irrelevant Silly Parentheses'.[2]

McCarthy invented LISP in the mid-1950s, but astonishingly, nearly 70 years later, it is still regularly taught and used across the world. (I use it every day.) Think about that for a moment: when McCarthy invented LISP, Dwight D. Eisenhower was President of the United States, Nikita Khrushchev was First Secretary of the Communist Party of the Soviet Union and in China Chairman Mao Zedong was overseeing the implementation of his first five-year plan. There were no more than a handful of computers in the whole world. And the programming language McCarthy invented then is still routinely used today.

Born to immigrant parents in Boston, McCarthy demonstrated an unusual aptitude for mathematics at an early age. After graduating in mathematics from Caltech (California Institute of Technology), he was appointed to an associate professorship at Dartmouth College in New Hampshire while he was still in his twenties. McCarthy had become interested in computing before joining Dartmouth, and in 1955 he submitted a proposal to the Rockefeller Institute in the hope of obtaining funds to organize a summer school at Dartmouth. If you are not an academic, the idea of 'summer schools' for adults may sound a little strange, but they are a well-established and fondly regarded academic tradition even today. The basic idea is to bring together a group of researchers with common interests from across the world, and give them the opportunity to meet and work together for an extended period. They are held in summer because, of course, teaching has finished for the year, and academics have a big chunk of time without lecturing commitments. Naturally, the goal is to organize the summer school in an attractive location, and a lively programme of social events is essential.

Another essential requirement for a memorable summer school is a star-studded delegate list. With the benefit of hindsight, we can see that the Dartmouth summer school brought together most of the key individuals that would define the field of AI for decades ahead. One name on the invitation list is particularly poignant. The Princeton-based mathematician John Forbes Nash Jr had gained his PhD in mathematics six years earlier, introducing (in a thesis just 28 pages long) a concept called 'non-cooperative games'. The ideas Nash introduced became cornerstones of economic theory in the decades that followed, and ultimately earned him a Nobel Prize in 1994. But Nash was unable to enjoy the recognition his work was attracting. Just a

few years after his PhD, Nash was consumed by episodes of paranoia and delusion, which took him out of academic life for decades. Happily, he recovered sufficiently that he was able to receive his Nobel Prize in 1994; his life was made the subject of the award-winning book and film *A Beautiful Mind.*[3]

The Dartmouth delegate list is also intriguing for another reason. Apart from the obvious presence of academics, the school hosted representatives from industry, government and the military (and even the RAND Corporation – the California-based thinktank made notorious in the 1960s for dispassionately debating how to 'win' a nuclear war). Only a decade earlier, the Manhattan Project had combined the capabilities of US academia, industry, government and military to develop the first atomic bomb – an unequivocal demonstration of US scientific and technological power. This combination – of academia, industry and the government/military – was characteristic of the US development of computer technology in the decades that followed the Second World War, and was central to the establishment of the US as the international leader in AI over the next six decades.

When McCarthy wrote his funding proposal for the Rockefeller Institute in 1955, he had to give a name to the event, and he chose 'artificial intelligence'. In what would become something of a weary tradition for AI, McCarthy had unrealistically high expectations for his event: 'We think that a significant advance can be made [...] if a carefully selected group of scientists work on it together for a summer.'[4]

By the end of the summer school, the delegates had made no real progress but McCarthy's chosen name had stuck, and thus was a new academic discipline formed.

Unfortunately, many have subsequently had occasion to regret McCarthy's choice of name for the field he founded. For one thing, *artificial* can be read as *fake*, or *ersatz* – and who wants *fake intelligence*? Moreover, the word *intelligence* suggests that *intellect* is key. In fact, many of the tasks that the AI community has worked so hard on since 1956 don't seem to require *intelligence* when people do them. On the contrary, as we saw in the previous chapter, many of the most important and difficult problems that AI has struggled with over the past 60 years don't seem to be *intellectual* at all – a fact that has repeatedly been the cause of consternation and confusion for those new to the field.

But artificial intelligence was the name that McCarthy chose, and the name that persists to this day. From McCarthy's summer school, there is an unbroken thread of research by way of the participants in the summer school and their academic descendants, right down to the present day. AI in its recognizably modern form began that summer, and the beginnings of AI seemed to be very promising indeed.

The period following the Dartmouth summer school was one of excitement and growth. And, for a while at least, it seemed like there was rapid progress. Four delegates of the summer school went on to dominate AI in the decades that followed. McCarthy himself founded the AI lab at Stanford University in the heart of what is now Silicon Valley; Marvin Minsky founded the AI lab at MIT in Cambridge, Massachusetts; Alan Newell and his PhD supervisor Herb Simon went to Carnegie Mellon University (CMU). These four individuals, and the AI systems that they and their students built, are totems for AI researchers of my generation.

But there was a good deal of naivety in the Golden Age, with researchers making reckless and grandiose predictions about the likely speed of progress in the field, which have haunted AI ever since. By the mid-1970s, the good times were over, and a vicious backlash began – an AI boom and bust cycle destined to be repeated over the coming decades. But, however critically history may judge this period, it is hard for me to contemplate the characters of this time, and the work they did, with anything other than affection.

Divide and Conquer

As we've seen, General AI is a large and very nebulous target – it is hard to approach directly. Instead, the strategy adopted during the Golden Age was that of *divide and conquer.* Thus, instead of starting out trying to build a complete general intelligent system, the approach adopted was to identify the various individual *capabilities* that seemed to be required for general-purpose AI, and to build systems that could demonstrate these capabilities. The implicit assumption was that, if we could succeed in building systems that demonstrate each of these individual capabilities, then, later on, assembling them into a whole would be straightforward. This fundamental assumption – that the way to progress towards General AI was to focus on the component capabilities of intelligent behaviour – became embedded as the standard methodology for AI research. There was a rush to build machines that could demonstrate these component capabilities.

So, what were the main capabilities that researchers focused on? The first, and as it turned out one of the most stubbornly difficult, is one that we take for granted: **perception**. A machine that is going to act intelligently in a particular environment needs to be able to get information about it. We perceive our world through various mechanisms, including the five senses: sight, sound, touch, smell and taste. So, one strand of research involved building **sensors** that provide analogues of these. Robots today use a wide range of artificial sensors to give them information about their environment – radars, infrared range-finders, ultrasonic range-finders, laser radar and so on. But *building* these sensors – which in itself is not trivial – is only part of the problem. However good the optics are on a digital camera, and no matter how many megapixels there are on the camera's image sensor, ultimately all that camera does is break down the image it is seeing into a grid, and then assign numbers to each cell in the grid, indicating colour and brightness. So, a robot equipped with the very best digital camera will, in the end, only receive a long list of numbers. The second challenge of perception is therefore to interpret those raw numbers: to understand what it is seeing. And this challenge turned out to be far, far harder than the problem of actually building the sensors.

Another key capability for general intelligent systems seems to be the ability to learn from experience, and this led to a strand of AI research called **machine learning**. Like the name 'artificial intelligence' itself, 'machine learning' is perhaps an unfortunate choice of terminology. It sounds like a machine somehow bootstrapping its own intelligence: starting from nothing and progressively getting smarter and smarter. In fact, machine learning is not like human learning: it is about learning from and making predictions about data. For example, one big success in machine learning over the past decade is in programs that can recognize faces in pictures. The way this is usually done involves providing a program with examples of the things that you are trying to learn. Thus, a program to recognize faces would be trained by giving it many pictures labelled with the names of the people that appear in the pictures. The goal is that, when subsequently presented with solely an image, the program would be able to correctly give the name of the person pictured.

Problem solving and **planning** are two related capabilities that also seem to be associated with intelligent behaviour. They both require being able to achieve a goal using a given repertoire of actions; the challenge is to find the right sequence of actions. Playing a board game such as chess or Go would be an example: the goal is to win the game; the actions are the possible moves; the challenge is to figure out which moves to make. As we will see, one of the most fundamental challenges in problem solving and planning is that while they appear easy to do *in principle*, by considering all the possible alternatives, this approach doesn't work in practice, because there are far too many alternatives for it to be feasible.

Index

The page references in this index correspond to the print edition from which this ebook was created, and clicking on them will take you to the location in the ebook where the equivalent print page would begin. To find a specific word or phrase from the index, please use the search feature of your ebook reader.

A

A* 77
À la recherche du temps perdu (Proust) 205–8
accountability 257
Advanced Research Projects Agency (ARPA) 87–8
adversarial machine learning 190
AF (Artificial Flight) parable 127–9, 243
agent-based AI 136–49
agent-based interfaces 147, 149
'Agents That Reduce Work and Information Overload' (Maes) 147–8
AGI (Artificial General Intelligence) 41
AI
– difficulty of 24–8
– ethical 246–62, 284, 285
– future of 7–8
– General 42, 53, 116, 119–20
– Golden Age of 47–88
– history of 5–7
– meaning of 2–4
– narrow 42
– origin of name 51–2
– strong 36–8, 41, 309–14
– symbolic 42–3, 44
– varieties of 36–8
– weak 36–8
AI winter 87–8
AI-complete problems 84
'Alchemy and AI' (Dreyfus) 85
AlexNet 187
algorithmic bias 287–9, 292–3
alienation 274–7
allocative harm 287–8
AlphaFold 214
AlphaGo 196–9
AlphaGo Zero 199
AlphaZero 199–200
Alvey programme 100

Amazon 275–6
Apple Watch 218
Argo AI 232
arithmetic 24–6
Arkin, Ron 284
ARPA (Advanced Research Projects Agency) 87–8
Artificial Flight (AF) parable 127–9, 243
Artificial General Intelligence (AGI) 41
artificial intelligence *see* AI
artificial languages 56
Asilomar principles 254–6
Asimov, Isaac 244–6
Atari 2600 games console 192–6, 327–8
augmented reality 296–7
automated diagnosis 220–1
automated translation 204–8
automation 265, 267–72
autonomous drones 282–4
Autonomous Vehicle Disengagement Reports 231
autonomous vehicles *see* driverless cars
autonomous weapons 281–7
autonomy levels 227–8
Autopilot 228–9

B

backprop/backpropagation 182–3
backward chaining 94
Bayes nets 158
Bayes' Theorem 155–8, 365–7
Bayesian networks 158
behavioural AI 132–7
beliefs 108–10
bias 172
black holes 213–14
Blade Runner 38
Blocks World 57–63, 126–7
blood diseases 94–8
board games 26, 75–6
Boole, George 107
brains 43, 306, 330–1 *see also* electronic brains
branching factors 73
Breakout (video game) 193–5
Brooks, Rodney 125–9, 132, 134, 243
bugs 258

C

Campaign to Stop Killer Robots 286
CaptionBot 201–4
Cardiogram 215
cars 27–8, 155, 223–35

certainty factors 97
ceteris paribus preferences 262
chain reactions 242–3
chatbots 36
checkers 75–7
chess 163–4, 199
Chinese room 311–14
choice under uncertainty 152–3
combinatorial explosion 74, 80–1
common values and norms 260
common-sense reasoning 121–3 *see also* reasoning
COMPAS 280
complexity barrier 77–85
comprehension 38–41
computational complexity 77–85
computational effort 129
computers
– decision making 23–4
– early developments 20
– as electronic brains 20–4
– intelligence 21–2
– programming 21–2
– reliability 23
– speed of 23
– tasks for 24–8
– unsolved problems 28
'Computing Machinery and Intelligence' (Turing) 32
confirmation bias 295
conscious machines 327–30
consciousness 305–10, 314–17, 331–4
consensus reality 296–8
consequentialist theories 249
contradictions 122–3
conventional warfare 286
credit assignment problem 173, 196
Criado Perez, Caroline 291–2
crime 277–81
Cruise Automation 232
curse of dimensionality 172
cutlery 261
Cybernetics (Wiener) 29
Cyc 114–21, 208

D

DARPA (Defense Advanced Research Projects Agency) 87–8, 225–6
Dartmouth summer school 1955 50–2
decidable problems 78–9
decision problems 15–19
deduction 106
deep learning 168, 184–90, 208
DeepBlue 163–4

PENGUIN BOOKS

UK | USA | Canada | Ireland | Australia
India | New Zealand | South Africa

Penguin Books is part of the Penguin Random House group of companies whose addresses can be found at global.penguinrandomhouse.com.

First published 2020

Cover by Matthew Young

Book design by Matthew Young

ISBN: 978-0-241-33391-4