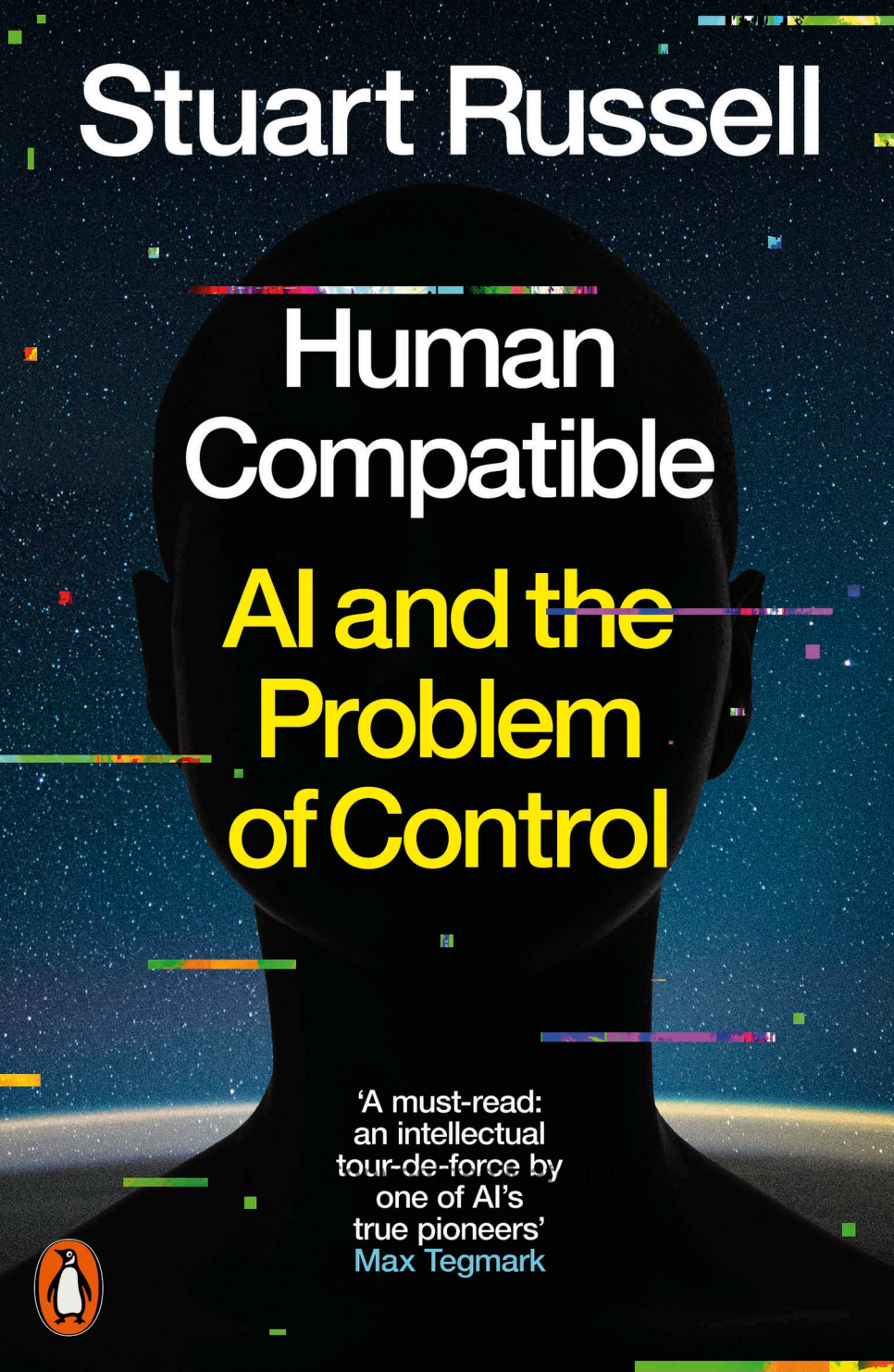

Stuart Russell
Human Compatible
AI and the Problem of Control
'A must-read: an intellectual tour-de-force by one of AI's true pioneers'
Max Tegmark

PENGUIN BOOKS

HUMAN COMPATIBLE

'In clear and compelling language, Stuart Russell describes the huge potential benefits of Artificial Intelligence, as well as the hazards and ethical challenges. It's especially welcome that a respected leading authority should offer this balanced appraisal, avoiding both hype and scaremongering'
Lord Rees, Astronomer Royal and author of *On the Future*

'A delicious excursion . . . Russell's exciting book goes deep, while sparkling with dry witticisms' *Wall Street Journal*

'The same mix of de-mystifying authority and practical advice that Dr Benjamin Spock once brought to the care and raising of children, Dr Stuart Russell now brings to the care, raising, and yes, disciplining of machines. He has written the book that most – but perhaps not all – machines would like you to read'
George Dyson, author of *Turing's Cathedral*

'Brilliantly clear and fascinating' Steven Poole, *Spectator*

'A thought-provoking and highly readable account of the past, present and future of AI . . . Russell deploys a bracing intellectual rigour . . . but a laconic style and dry humour keep his book accessible to the lay reader' *Financial Times*

'Russell is an assiduous and conscientious scholar . . . he provides a wealth of information. This is one of those intellectual voyages where both the journey and the destination matter'
John Naughton, *Literary Review*

'It's asking a lot of a book about the potential end of civilization to be strewn with humour and wry asides, but this is what Russell manages. He tackles what he believes to be a fundamental flaw at the heart of artificial intelligence, one that could spell disaster if left unfixed'
Ian Sample, *Guardian*, Books of the Year

'This beautifully written book addresses a fundamental challenge for humanity: increasingly intelligent machines that do what we ask but not what we really intend. Essential reading if you care about our future' Yoshua Bengio, winner of the 2019 Turing Award and co-author of *Deep Learning*

'I just finished Stuart Russell's marvellous book on AI safety, *Human Compatible*, and I can't recommend it highly enough!'
Tim O'Reilly, author of *WTF?: What's the Future and Why It's Up to Us*

'*Human Compatible* made me a convert to Russell's concerns with our ability to control our upcoming creation – super-intelligent machines. Unlike outside alarmists and futurists, Russell is a leading authority on AI. His new book will educate the public about AI more than any book I can think of, and is a delightful and uplifting read' Judea Pearl, author of *The Book of Why*

'Stuart Russell, one of the most important AI scientists of the last 25 years, may have written the most important book about AI so far, on one of the most important questions of the 21st century' James Manyika, Chairman and director of McKinsey Global Institute

'No researcher has argued more persuasively about the risks of AI, nor has shown more clearly a pathway forward. Anyone who takes the future seriously should pay attention' Brian Christian, author of *Algorithms to Live By*

'Can we coexist happily with the intelligent machines that humans will create? "Yes," answers *Human Compatible*, "but first . . .". Through a brilliant reimagining of the foundations of artificial intelligence, Russell takes you on a journey from the very beginning, explaining the questions raised by an AI-driven society and beautifully making the case for how to ensure machines remain beneficial to humans' Tabitha Goldstaub, co-founder of CognitionX and Head of the UK Government's AI Council

'The wealth of knowledge of a prominent AI researcher and the persuasive clarity and wit of a brilliant educator' Jaan Tallinn, co-founder of Skype

ABOUT THE AUTHOR

Stuart Russell is a professor of Computer Science and holder of the Smith-Zadeh Chair in Engineering at the University of California, Berkeley, and an Honorary Fellow of Wadham College, University of Oxford. He has advised Number 10 and the United Nations about the risks of AI. In 1990 he received the Presidential Young Investigator Award of the National Science Foundation, in 1995 he was co-winner of the Computers and Thought Award, and in 2005 he received the ACM Karlstrom Outstanding Educator Award. He is the author (with Peter Norvig) of *Artificial Intelligence: A Modern Approach*, the Number One bestselling textbook in AI which is used in over 1,300 universities in 118 countries around the world. He was born in England and lives in Berkeley.

Human Compatible

Artificial Intelligence and the Problem of Control

Stuart Russell

PENGUIN BOOKS

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 in the USA by Viking 2019
First published in Great Britain by Allen Lane 2019
Published in Penguin Books 2020
Reprinted with an Afterword, 2023
001

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

The authorized representative in the EEA is Penguin Random House Ireland,
Morrison Chambers, 32 Nassau Street, Dublin D02 YH68

A CIP catalogue record for this book is available from the British Library

ISBN: 978–0–141–98750–7

www.greenpenguin.co.uk

Penguin Random House is committed to a
sustainable future for our business, our readers
and our planet. This book is made from Forest
Stewardship Council®-certified paper.

For Loy, Gordon, Lucy, George, and Isaac

CONTENTS

PREFACE

Why This Book? Why Now?

This book is about the past, present, and future of our attempt to understand and create intelligence. This matters, not because AI is rapidly becoming a pervasive aspect of the present but because it is the dominant technology of the future. The world's great powers are waking up to this fact, and the world's largest corporations have known it for some time. We cannot predict exactly how the technology will develop or on what timeline. Nevertheless, we must plan for the possibility that machines will far exceed the human capacity for decision making in the real world. What then?

Everything civilization has to offer is the product of our intelligence; gaining access to considerably greater intelligence would be the biggest event in human history. The purpose of the book is to explain why it might be the last event in human history and how to make sure that it is not.

Overview of the Book

The book has three parts. The first part (Chapters 1 to 3) explores the idea of intelligence in humans and in machines. The material requires no technical background, but for those who are interested, it is supplemented by four appendices that explain some of the core concepts underlying present-day AI systems. The second part (Chapters 4 to 6) discusses some problems arising from imbuing machines with intelligence. I focus in particular on the problem of control: retaining absolute power over machines that are more powerful than us. The third part (Chapters 7 to 10) suggests a new way to think about AI and to ensure that machines remain beneficial to humans, forever. This Re-issue includes an Afterword covering developments between 2019 (the original publication date) and 2023.

The book is intended for a general audience but will, I hope, be of value in convincing specialists in artificial intelligence to rethink their fundamental assumptions.

1

IF WE SUCCEED

A long time ago, my parents lived in Birmingham, England, in a house near the university. They decided to move out of the city and sold the house to David Lodge, a professor of English literature. Lodge was by that time already a well-known novelist. I never met him, but I decided to read some of his books: C*hanging Places* and *Small World*. Among the principal characters were fictional academics moving from a fictional version of Birmingham to a fictional version of Berkeley, California. As I was an actual academic from the actual Birmingham who had just moved to the actual Berkeley, it seemed that someone in the Department of Coincidences was telling me to pay attention.

One particular scene from *Small World* struck me: The protagonist, an aspiring literary theorist, attends a major international conference and asks a panel of leading figures, "What follows if everyone agrees with you?" The question causes consternation, because the panelists had been more concerned with intellectual combat than ascertaining truth or attaining understanding. It occurred to me then that an analogous question could be asked of the leading figures in AI: "What if you succeed?" The field's goal had always been to create

human-level or superhuman AI, but there was little or no consideration of what would happen if we did.

A few years later, Peter Norvig and I began work on a new AI textbook, whose first edition appeared in 1995.[1] The book's final section is titled "What If We Do Succeed?" The section points to the possibility of good and bad outcomes but reaches no firm conclusions. By the time of the third edition in 2010, many people had finally begun to consider the possibility that superhuman AI might not be a good thing—but these people were mostly outsiders rather than mainstream AI researchers. By 2013, I became convinced that the issue not only belonged in the mainstream but was possibly the most important question facing humanity.

In November 2013, I gave a talk at the Dulwich Picture Gallery, a venerable art museum in south London. The audience consisted mostly of retired people—nonscientists with a general interest in intellectual matters—so I had to give a completely nontechnical talk. It seemed an appropriate venue to try out my ideas in public for the first time. After explaining what AI was about, I nominated five candidates for "biggest event in the future of humanity":

1. We all die (asteroid impact, climate catastrophe, pandemic, etc.).
2. We all live forever (medical solution to aging).
3. We invent faster-than-light travel and conquer the universe.
4. We are visited by a superior alien civilization.
5. We invent superintelligent AI.

I suggested that the fifth candidate, superintelligent AI, would be the winner, because it would help us avoid physical catastrophes and achieve eternal life and faster-than-light travel, if those were indeed possible. It would represent a huge leap—a discontinuity—in our civilization. The arrival of superintelligent AI is in many ways analogous to the arrival of a superior alien civilization but much more likely to

occur. Perhaps most important, AI, unlike aliens, is something over which we have some say.

Then I asked the audience to imagine what would happen if we received notice from a superior alien civilization that they would arrive on Earth in thirty to fifty years. The word *pandemonium* doesn't begin to describe it. Yet our response to the anticipated arrival of superintelligent AI has been . . . well, underwhelming begins to describe it. (In a later talk, I illustrated this in the form of the email exchange shown in figure 1.) Finally, I explained the significance of superintelligent AI as follows: "Success would be the biggest event in human history . . . and perhaps the last event in human history."

From: Superior Alien Civilization <sac12@sirius.canismajor.u>

To: humanity@UN.org

Subject: Contact

Be warned: we shall arrive in 30–50 years

From: humanity@UN.org

To: Superior Alien Civilization <sac12@sirius.canismajor.u>

Subject: Out of office: Re: Contact

Humanity is currently out of the office. We will respond to your message when we return. ☺

FIGURE 1: Probably not the email exchange that would follow the first contact by a superior alien civilization.

A few months later, in April 2014, I was at a conference in Iceland and got a call from National Public Radio asking if they could interview me about the movie *Transcendence*, which had just been released in the United States. Although I had read the plot summaries and reviews, I hadn't seen it because I was living in Paris at the time, and it would not be released there until June. It so happened, however, that

I had just added a detour to Boston on the way home from Iceland, so that I could participate in a Defense Department meeting. So, after arriving at Boston's Logan Airport, I took a taxi to the nearest theater showing the movie. I sat in the second row and watched as a Berkeley AI professor, played by Johnny Depp, was gunned down by anti-AI activists worried about, yes, superintelligent AI. Involuntarily, I shrank down in my seat. (Another call from the Department of Coincidences?) Before Johnny Depp's character dies, his mind is uploaded to a quantum supercomputer and quickly outruns human capabilities, threatening to take over the world.

On April 19, 2014, a review of *Transcendence*, co-authored with physicists Max Tegmark, Frank Wilczek, and Stephen Hawking, appeared in the *Huffington Post*. It included the sentence from my Dulwich talk about the biggest event in human history. From then on, I would be publicly committed to the view that my own field of research posed a potential risk to my own species.

How Did We Get Here?

The roots of AI stretch far back into antiquity, but its "official" beginning was in 1956. Two young mathematicians, John McCarthy and Marvin Minsky, had persuaded Claude Shannon, already famous as the inventor of information theory, and Nathaniel Rochester, the designer of IBM's first commercial computer, to join them in organizing a summer program at Dartmouth College. The goal was stated as follows:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think

> that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

Needless to say, it took much longer than a summer: we are still working on all these problems.

In the first decade or so after the Dartmouth meeting, AI had several major successes, including Alan Robinson's algorithm for general-purpose logical reasoning[2] and Arthur Samuel's checker-playing program, which taught itself to beat its creator.[3] The first AI bubble burst in the late 1960s, when early efforts at machine learning and machine translation failed to live up to expectations. A report commissioned by the UK government in 1973 concluded, "In no part of the field have the discoveries made so far produced the major impact that was then promised."[4] In other words, the machines just weren't smart enough.

My eleven-year-old self was, fortunately, unaware of this report. Two years later, when I was given a Sinclair Cambridge Programmable calculator, I just wanted to make it intelligent. With a maximum program size of thirty-six keystrokes, however, the Sinclair was not quite big enough for human-level AI. Undeterred, I gained access to the giant CDC 6600 supercomputer[5] at Imperial College London and wrote a chess program—a stack of punched cards two feet high. It wasn't very good, but it didn't matter. I knew what I wanted to do.

By the mid-1980s, I had become a professor at Berkeley, and AI was experiencing a huge revival thanks to the commercial potential of so-called expert systems. The second AI bubble burst when these systems proved to be inadequate for many of the tasks to which they were applied. Again, the machines just weren't smart enough. An AI winter ensued. My own AI course at Berkeley, currently bursting with over nine hundred students, had just twenty-five students in 1990.

The AI community learned its lesson: smarter, obviously, was better, but we would have to do our homework to make that happen. The

field became far more mathematical. Connections were made to the long-established disciplines of probability, statistics, and control theory. The seeds of today's progress were sown during that AI winter, including early work on large-scale probabilistic reasoning systems and what later became known as *deep learning*.

Beginning around 2011, deep learning techniques began to produce dramatic advances in speech recognition, visual object recognition, and machine translation—three of the most important open problems in the field. By some measures, machines now match or exceed human capabilities in these areas. In 2016 and 2017, DeepMind's AlphaGo defeated Lee Sedol, former world Go champion, and Ke Jie, the current champion—events that some experts predicted wouldn't happen until 2097, if ever.[6]

Now AI generates front-page media coverage almost every day. Thousands of start-up companies have been created, fueled by a flood of venture funding. Millions of students have taken online AI and machine learning courses, and experts in the area command salaries in the millions of dollars. Investments flowing from venture funds, national governments, and major corporations are in the tens of billions of dollars annually—more money in the last five years than in the entire previous history of the field. Advances that are already in the pipeline, such as self-driving cars and intelligent personal assistants, are likely to have a substantial impact on the world over the next decade or so. The potential economic and social benefits of AI are vast, creating enormous momentum in the AI research enterprise.

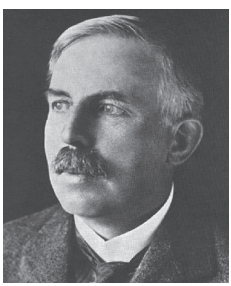
What Happens Next?

Does this rapid rate of progress mean that we are about to be overtaken by machines? No. There are several breakthroughs that have to happen before we have anything resembling machines with superhuman intelligence.

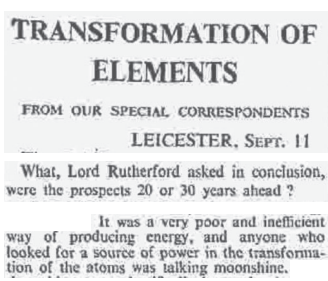
Scientific breakthroughs are notoriously hard to predict. To get a sense of just how hard, we can look back at the history of another field with civilization-ending potential: nuclear physics.

In the early years of the twentieth century, perhaps no nuclear physicist was more distinguished than Ernest Rutherford, the discoverer of the proton and the "man who split the atom" (figure 2[a]). Like his colleagues, Rutherford had long been aware that atomic nuclei stored immense amounts of energy; yet the prevailing view was that tapping this source of energy was impossible.

On September 11, 1933, the British Association for the Advancement of Science held its annual meeting in Leicester. Lord Rutherford addressed the evening session. As he had done several times before, he poured cold water on the prospects for atomic energy: "Anyone who looks for a source of power in the transformation of the atoms is talking moonshine." Rutherford's speech was reported in the *Times* of London the next morning (figure 2[b]).



(a)



TRANSFORMATION OF ELEMENTS

FROM OUR SPECIAL CORRESPONDENTS

LEICESTER, SEPT. 11

What, Lord Rutherford asked in conclusion, were the prospects 20 or 30 years ahead ?

It was a very poor and inefficient way of producing energy, and anyone who looked for a source of power in the transformation of the atoms was talking moonshine.

(b)

(c)

FIGURE 2: (a) Lord Rutherford, nuclear physicist. (b) Excerpts from a report in the *Times* of September 12, 1933, concerning a speech given by Rutherford the previous evening. (c) Leo Szilard, nuclear physicist.

Leo Szilard (figure 2[c]), a Hungarian physicist who had recently fled from Nazi Germany, was staying at the Imperial Hotel on Russell

Square in London. He read the *Times*' report at breakfast. Mulling over what he had read, he went for a walk and invented the neutron-induced nuclear chain reaction.[7] The problem of liberating nuclear energy went from impossible to essentially solved in less than twenty-four hours. Szilard filed a secret patent for a nuclear reactor the following year. The first patent for a nuclear weapon was issued in France in 1939.

The moral of this story is that betting against human ingenuity is foolhardy, particularly when our future is at stake. Within the AI community, a kind of denialism is emerging, even going as far as denying the possibility of success in achieving the long-term goals of AI. It's as if a bus driver, with all of humanity as passengers, said, "Yes, I am driving as hard as I can towards a cliff, but trust me, we'll run out of gas before we get there!"

I am not saying that success in AI will *necessarily* happen, and I think it's quite unlikely that it will happen in the next few years. It seems prudent, nonetheless, to prepare for the eventuality. If all goes well, it would herald a golden age for humanity, but we have to face the fact that we are planning to make entities that are far more powerful than humans. How do we ensure that they never, ever have power over us?

To get just an inkling of the fire we're playing with, consider how content-selection algorithms function on social media. They aren't particularly intelligent, but they are in a position to affect the entire world because they directly influence billions of people. Typically, such algorithms are designed to maximize *click-through*, that is, the probability that the user clicks on presented items. The solution is simply to present items that the user likes to click on, right? Wrong. The solution is to change the user's preferences so that they become more predictable. A more predictable user can be fed items that they are likely to click on, thereby generating more revenue. People with more extreme political views tend to be more predictable in which items they will click on. (Possibly there is a category of articles that

die-hard centrists are likely to click on, but it's not easy to imagine what this category consists of.) Like any rational entity, the algorithm learns how to modify the state of its environment—in this case, the user's mind—in order to maximize its own reward.[8] The consequences include the resurgence of fascism, the dissolution of the social contract that underpins democracies around the world, and potentially the end of the European Union and NATO. Not bad for a few lines of code, even if it had a helping hand from some humans. Now imagine what a *really* intelligent algorithm would be able to do.

What Went Wrong?

The history of AI has been driven by a single mantra: "The more intelligent the better." I am convinced that this is a mistake—not because of some vague fear of being superseded but because of the way we have understood intelligence itself.

The concept of intelligence is central to who we are—that's why we call ourselves *Homo sapiens*, or "wise man." After more than two thousand years of self-examination, we have arrived at a characterization of intelligence that can be boiled down to this:

> *Humans are intelligent to the extent that our actions can be expected to achieve our objectives.*

All those other characteristics of intelligence—perceiving, thinking, learning, inventing, and so on—can be understood through their contributions to our ability to act successfully. From the very beginnings of AI, intelligence in machines has been defined in the same way:

> *Machines are intelligent to the extent that their actions can be expected to achieve their objectives.*

Because machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build optimizing machines, we feed objectives into them, and off they go.

This general approach is not unique to AI. It recurs throughout the technological and mathematical underpinnings of our society. In the field of control theory, which designs control systems for everything from jumbo jets to insulin pumps, the job of the system is to minimize a *cost function* that typically measures some deviation from a desired behavior. In the field of economics, mechanisms and policies are designed to maximize the *utility* of individuals, the *welfare* of groups, and the *profit* of corporations.[9] In operations research, which solves complex logistical and manufacturing problems, a solution maximizes an expected *sum of rewards* over time. Finally, in statistics, learning algorithms are designed to minimize an expected *loss function* that defines the cost of making prediction errors.

Evidently, this general scheme—which I will call the *standard model*—is widespread and extremely powerful. Unfortunately, *we don't want machines that are intelligent in this sense.*

The drawback of the standard model was pointed out in 1960 by Norbert Wiener, a legendary professor at MIT and one of the leading mathematicians of the mid-twentieth century. Wiener had just seen Arthur Samuel's checker-playing program learn to play checkers far better than its creator. That experience led him to write a prescient but little-known paper, "Some Moral and Technical Consequences of Automation."[10] Here's how he states the main point:

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively . . . we had better be quite sure that the purpose put into the machine is the purpose which we really desire.

"The purpose put into the machine" is exactly the objective that machines are optimizing in the standard model. If we put the wrong

objective into a machine that is more intelligent than us, it will achieve the objective, and we lose. The social-media meltdown I described earlier is just a foretaste of this, resulting from optimizing the wrong objective on a global scale with fairly unintelligent algorithms. In Chapter 5, I spell out some far worse outcomes.

All this should come as no great surprise. For thousands of years, we have known the perils of getting exactly what you wish for. In every story where someone is granted three wishes, the third wish is always to undo the first two wishes.

In summary, it seems that the march towards superhuman intelligence is unstoppable, but success might be the undoing of the human race. Not all is lost, however. We have to understand where we went wrong and then fix it.

Can We Fix It?

The problem is right there in the basic definition of AI. We say that machines are intelligent to the extent that their actions can be expected to achieve *their* objectives, but we have no reliable way to make sure that *their* objectives are the same as *our* objectives.

What if, instead of allowing machines to pursue *their* objectives, we insist that they pursue *our* objectives? Such a machine, if it could be designed, would be not just *intelligent* but also *beneficial* to humans. So let's try this:

> *Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives.*

This is probably what we should have done all along.

The difficult part, of course, is that our objectives are in us (all eight billion of us, in all our glorious variety) and not in the machines. It is, nonetheless, possible to build machines that are beneficial in

exactly this sense. Inevitably, these machines will be uncertain about our objectives—after all, we are uncertain about them ourselves—but it turns out that this is a feature, not a bug (that is, a good thing and not a bad thing). Uncertainty about objectives implies that machines will necessarily defer to humans: they will ask permission, they will accept correction, and they will allow themselves to be switched off.

Removing the assumption that machines should have a definite objective means that we will need to tear out and replace part of the foundations of artificial intelligence—the basic definitions of what we are trying to do. That also means rebuilding a great deal of the superstructure—the accumulation of ideas and methods for actually doing AI. The result will be a new relationship between humans and machines, one that I hope will enable us to navigate the next few decades successfully.

2

INTELLIGENCE IN HUMANS AND MACHINES

When you arrive at a dead end, it's a good idea to retrace your steps and work out where you took a wrong turn. I have argued that the standard model of AI, wherein machines optimize a fixed objective supplied by humans, is a dead end. The problem is not that we might *fail* to do a good job of building AI systems; it's that we might *succeed* too well. The very definition of success in AI is wrong.

So let's retrace our steps, all the way to the beginning. Let's try to understand how our concept of intelligence came about and how it came to be applied to machines. Then we have a chance of coming up with a better definition of what counts as a good AI system.

Intelligence

How does the universe work? How did life begin? Where are my keys? These are fundamental questions worthy of thought. But who is asking these questions? How am I answering them? How can a handful

of matter—the few pounds of pinkish-gray blancmange we call a brain—perceive, understand, predict, and manipulate a world of unimaginable vastness? Before long, the mind turns to examine itself.

We have been trying for thousands of years to understand how our minds work. Initially, the purposes included curiosity, self-management, persuasion, and the rather pragmatic goal of analyzing mathematical arguments. Yet every step towards an explanation of how the mind works is also a step towards the creation of the mind's capabilities in an artifact—that is, a step towards artificial intelligence.

Before we can understand how to create intelligence, it helps to understand what it is. The answer is not to be found in IQ tests, or even in Turing tests, but in a simple relationship between what we perceive, what we want, and what we do. Roughly speaking, an entity is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived.

Evolutionary origins

Consider a lowly bacterium, such as *E. coli*. It is equipped with about half a dozen flagella—long, hairlike tentacles that rotate at the base either clockwise or counterclockwise. (The rotary motor itself is an amazing thing, but that's another story.) As *E. coli* floats about in its liquid home—your lower intestine—it alternates between rotating its flagella clockwise, causing it to "tumble" in place, and counterclockwise, causing the flagella to twine together into a kind of propeller so the bacterium swims in a straight line. Thus, *E. coli* does a sort of random walk—swim, tumble, swim, tumble—that allows it to find and consume glucose rather than staying put and dying of starvation.

If this were the whole story, we wouldn't say that *E. coli* is particularly intelligent, because its actions would not depend in any way on its environment. It wouldn't be making any decisions, just executing a fixed behavior that evolution has built into its genes. But this isn't the whole story. When *E. coli* senses an increasing concentration of

glucose, it swims longer and tumbles less, and it does the opposite when it senses a decreasing concentration of glucose. So, what it does (swim towards glucose) is likely to achieve what it wants (more glucose, let's assume), given what it has perceived (an increasing glucose concentration).

Perhaps you are thinking, "But evolution built this into its genes too! How does that make it intelligent?" This is a dangerous line of reasoning, because evolution built the basic design of your brain into your genes too, and presumably you wouldn't wish to deny your own intelligence on that basis. The point is that what evolution has built into *E. coli*'s genes, as it has into yours, is a mechanism whereby the bacterium's behavior varies according to what it perceives in its environment. Evolution doesn't know, in advance, where the glucose is going to be or where your keys are, so putting the capability to find them into the organism is the next best thing.

Now, *E. coli* is no intellectual giant. As far as we know, it doesn't remember where it has been, so if it goes from A to B and finds no glucose, it's just as likely to go back to A. If we construct an environment where every attractive glucose gradient leads only to a spot of phenol (which is a poison for *E. coli*), the bacterium will keep following those gradients. It never learns. It has no brain, just a few simple chemical reactions to do the job.

A big step forward occurred with *action potentials*, which are a form of electrical signaling that first evolved in single-celled organisms around a billion years ago. Later multicellular organisms evolved specialized cells called *neurons* that use electrical action potentials to carry signals rapidly—up to 120 meters per second, or 270 miles per hour—within the organism. The connections between neurons are called *synapses*. The strength of the synaptic connection dictates how much electrical excitation passes from one neuron to another. By changing the strength of synaptic connections, animals learn.[1] Learning confers a huge evolutionary advantage, because the animal can adapt to a range of circumstances. Learning also speeds up the rate of evolution itself.

Initially, neurons were organized into *nerve nets*, which are distributed throughout the organism and serve to coordinate activities such as eating and digestion or the timed contraction of muscle cells across a wide area. The graceful propulsion of jellyfish is the result of a nerve net. Jellyfish have no brains at all.

Brains came later, along with complex sense organs such as eyes and ears. Several hundred million years after jellyfish emerged with their nerve nets, we humans arrived with our big brains—a hundred billion (10^{11}) neurons and a quadrillion (10^{15}) synapses. While slow compared to electronic circuits, the "cycle time" of a few milliseconds per state change is fast compared to most biological processes. The human brain is often described by its owners as "the most complex object in the universe," which probably isn't true but is a good excuse for the fact that we still understand little about how it really works. While we know a great deal about the biochemistry of neurons and synapses and the anatomical structures of the brain, the neural implementation of the *cognitive* level—learning, knowing, remembering, reasoning, planning, deciding, and so on—is still mostly anyone's guess.[2] (Perhaps that will change as we understand more about AI, or as we develop ever more precise tools for measuring brain activity.) So, when one reads in the media that such-and-such AI technique "works just like the human brain," one may suspect it's either just someone's guess or plain fiction.

In the area of *consciousness*, we really do know nothing, so I'm going to say nothing. No one in AI is working on making machines conscious, nor would anyone know where to start, and no behavior has consciousness as a prerequisite. Suppose I give you a program and ask, "Does this present a threat to humanity?" You analyze the code and indeed, when run, the code will form and carry out a plan whose result will be the destruction of the human race, just as a chess program will form and carry out a plan whose result will be the defeat of any human who faces it. Now suppose I tell you that the code, when run, also creates a form of machine consciousness. Will that change your

prediction? Not at all. It makes *absolutely no difference*.[3] Your prediction about its behavior is exactly the same, because the prediction is based on the code. All those Hollywood plots about machines mysteriously becoming conscious and hating humans are really missing the point: it's competence, not consciousness, that matters.

There is one important cognitive aspect of the brain that we *are* beginning to understand—namely, the *reward system*. This is an internal signaling system, mediated by dopamine, that connects positive and negative stimuli to behavior. Its workings were discovered by the Swedish neuroscientist Nils-Åke Hillarp and his collaborators in the late 1950s. It causes us to seek out positive stimuli, such as sweet-tasting foods, that increase dopamine levels; it makes us avoid negative stimuli, such as hunger and pain, that decrease dopamine levels. In a sense it's quite similar to *E. coli*'s glucose-seeking mechanism, but much more complex. It comes with built-in methods for learning, so that our behavior becomes more effective at obtaining reward over time. It also allows for delayed gratification, so that we learn to desire things such as money that provide eventual reward rather than immediate reward. One reason we understand the brain's reward system is that it resembles the method of *reinforcement learning* developed in AI, for which we have a very solid theory.[4]

From an evolutionary point of view, we can think of the brain's reward system, just like *E. coli*'s glucose-seeking mechanism, as a way of improving evolutionary fitness. Organisms that are more effective in seeking reward—that is, finding delicious food, avoiding pain, engaging in sexual activity, and so on—are more likely to propagate their genes. It is extraordinarily difficult for an organism to decide what actions are most likely, in the long run, to result in successful propagation of its genes, so evolution has made it easier for us by providing built-in signposts.

These signposts are not perfect, however. There are ways to obtain reward that probably *reduce* the likelihood that one's genes will propagate. For example, taking drugs, drinking vast quantities of sugary

carbonated beverages, and playing video games for eighteen hours a day all seem counterproductive in the reproduction stakes. Moreover, if you were given direct electrical access to your reward system, you would probably self-stimulate without stopping until you died.[5]

The misalignment of reward signals and evolutionary fitness doesn't affect only isolated individuals. On a small island off the coast of Panama lives the pygmy three-toed sloth, which appears to be addicted to a Valium-like substance in its diet of red mangrove leaves and may be going extinct.[6] Thus, it seems that an entire species can disappear if it finds an ecological niche where it can satisfy its reward system in a maladaptive way.

Barring these kinds of accidental failures, however, learning to maximize reward in natural environments will usually improve one's chances for propagating one's genes and for surviving environmental changes.

Evolutionary accelerator

Learning is good for more than surviving and prospering. It also *speeds up evolution*. How could this be? After all, learning doesn't change one's DNA, and evolution is all about changing DNA over generations. The connection between learning and evolution was proposed in 1896 by the American psychologist James Baldwin[7] and independently by the British ethologist Conwy Lloyd Morgan[8] but not generally accepted at the time.

The Baldwin effect, as it is now known, can be understood by imagining that evolution has a choice between creating an *instinctive* organism whose every response is fixed in advance and creating an *adaptive* organism that learns what actions to take. Now suppose, for the purposes of illustration, that the optimal instinctive organism can be coded as a six-digit number, say, 472116, while in the case of the adaptive organism, evolution specifies only 472*** and the organism itself has to fill in the last three digits by learning during its lifetime.

Clearly, if evolution has to worry about choosing only the first three digits, its job is much easier; the adaptive organism, in learning the last three digits, is doing in one lifetime what evolution would have taken many generations to do. So, provided the adaptive organisms can survive while learning, it seems that the capability for learning constitutes an evolutionary shortcut. Computational simulations suggest that the Baldwin effect is real.[9] The effects of culture only accelerate the process, because an organized civilization protects the individual organism while it is learning and passes on information that the individual would otherwise need to learn for itself.

The story of the Baldwin effect is fascinating but incomplete: it assumes that learning and evolution necessarily point in the same direction. That is, it assumes that whatever internal feedback signal defines the direction of learning within the organism is perfectly aligned with evolutionary fitness. As we have seen in the case of the pygmy three-toed sloth, this does not seem to be true. At best, built-in mechanisms for learning provide only a crude hint of the long-term consequences of any given action for evolutionary fitness. Moreover, one has to ask, "How did the reward system get there in the first place?" The answer, of course, is by an evolutionary process, one that internalized a feedback mechanism that is at least somewhat aligned with evolutionary fitness.[10] Clearly, a learning mechanism that caused organisms to run away from potential mates and towards predators would not last long.

Thus, we have the Baldwin effect to thank for the fact that neurons, with their capabilities for learning and problem solving, are so widespread in the animal kingdom. At the same time, it is important to understand that evolution doesn't really care whether you have a brain or think interesting thoughts. Evolution considers you only as an *agent*, that is, something that acts. Such worthy intellectual characteristics as logical reasoning, purposeful planning, wisdom, wit, imagination, and creativity may be essential for making an agent intelligent, or they may not. One reason artificial intelligence is so fascinating is that

it offers a potential route to understanding these issues: we may come to understand both how these intellectual characteristics make intelligent behavior possible and why it's impossible to produce truly intelligent behavior without them.

Rationality for one

From the earliest beginnings of ancient Greek philosophy, the concept of intelligence has been tied to the ability to perceive, to reason, and to act *successfully*.[11] Over the centuries, the concept has become both broader in its applicability and more precise in its definition.

Aristotle, among others, studied the notion of successful reasoning—methods of logical deduction that would lead to true conclusions given true premises. He also studied the process of deciding how to act—sometimes called *practical reasoning*—and proposed that it involved deducing that a certain course of action would achieve a desired goal:

> We deliberate not about ends, but about means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade. . . . They assume the end and consider how and by what means it is attained, and if it seems easily and best produced thereby; while if it is achieved by one means only they consider *how* it will be achieved by this and by what means *this* will be achieved, till they come to the first cause . . . and what is last in the order of analysis seems to be first in the order of becoming. And if we come on an impossibility, we give up the search, e.g., if we need money and this cannot be got; but if a thing appears possible we try to do it.[12]

This passage, one might argue, set the tone for the next two-thousand-odd years of Western thought about rationality. It says that the "end"—what the person wants—is fixed and given; and it says that the rational

action is one that, according to logical deduction across a sequence of actions, "easily and best" produces the end.

Aristotle's proposal seems reasonable, but it isn't a complete guide to rational behavior. In particular, it omits the issue of uncertainty. In the real world, reality has a tendency to intervene, and few actions or sequences of actions are truly guaranteed to achieve the intended end. For example, it is a rainy Sunday in Paris as I write this sentence, and on Tuesday at 2:15 p.m. my flight to Rome leaves from Charles de Gaulle Airport, about forty-five minutes from my house. I plan to leave for the airport around 11:30 a.m., which should give me plenty of time, but it probably means at least an hour sitting in the departure area. Am I *certain* to catch the flight? Not at all. There could be huge traffic jams, the taxi drivers may be on strike, the taxi I'm in may break down or the driver may be arrested after a high-speed chase, and so on. Instead, I could leave for the airport on Monday, a whole day in advance. This would greatly reduce the chance of missing the flight, but the prospect of a night in the departure lounge is not an appealing one. In other words, my plan involves a *trade-off* between the certainty of success and the cost of ensuring that degree of certainty. The following plan for buying a house involves a similar trade-off: buy a lottery ticket, win a million dollars, then buy the house. This plan "easily and best" produces the end, but it's not very likely to succeed. The difference between this harebrained house-buying plan and my sober and sensible airport plan is, however, just a matter of degree. Both are gambles, but one seems more rational than the other.

It turns out that gambling played a central role in generalizing Aristotle's proposal to account for uncertainty. In the 1560s, the Italian mathematician Gerolamo Cardano developed the first mathematically precise theory of probability—using dice games as his main example. (Unfortunately, his work was not published until 1663.[13]) In the seventeenth century, French thinkers including Antoine Arnauld and Blaise Pascal began—for assuredly mathematical reasons—to

study the question of rational decisions in gambling.[14] Consider the following two bets:

A: 20 percent chance of winning $10
B: 5 percent chance of winning $100

The proposal the mathematicians came up with is probably the same one you would come up with: compare the *expected values* of the bets, which means the average amount you would expect to get from each bet. For bet A, the expected value is 20 percent of $10, or $2. For bet B, the expected value is 5 percent of $100, or $5. So bet B is better, according to this theory. The theory makes sense, because if the same bets are offered over and over again, a bettor who follows the rule ends up with more money than one who doesn't.

In the eighteenth century, the Swiss mathematician Daniel Bernoulli noticed that this rule didn't seem to work well for larger amounts of money.[15] For example, consider the following two bets:

A: 100 percent chance of getting $10,000,000
(expected value $10,000,000)
B: 1 percent chance of getting $1,000,000,100
(expected value $10,000,001)

Most readers of this book, as well as its author, would prefer bet A to bet B, even though the expected-value rule says the opposite! Bernoulli posited that bets are evaluated not according to expected monetary value but according to expected *utility*. Utility—the property of being useful or beneficial to a person—was, he suggested, an internal, subjective quantity related to, but distinct from, monetary value. In particular, utility exhibits *diminishing returns with respect to money*. This means that the utility of a given amount of money is not strictly proportional to the amount but grows more slowly. For example, the utility of having $1,000,000,100 is much less than a hundred times

the utility of having $10,000,000. How much less? You can ask yourself! What would the odds of winning a billion dollars have to be for you to give up a guaranteed ten million? I asked this question of the graduate students in my class and their answer was around 50 percent, meaning that bet B would have an expected value of $500 million to match the desirability of bet A. Let me say that again: bet B would have an expected dollar value fifty times greater than bet A, but the two bets would have equal utility.

Bernoulli's introduction of utility—an invisible property—to explain human behavior via a mathematical theory was an utterly remarkable proposal for its time. It was all the more remarkable for the fact that, unlike monetary amounts, the utility values of various bets and prizes are not directly observable; instead, utilities are to be *inferred* from the *preferences* exhibited by an individual. It would be two centuries before the implications of the idea were fully worked out and it became broadly accepted by statisticians and economists.

In the middle of the twentieth century, John von Neumann (a great mathematician after whom the standard "von Neumann architecture" for computers was named[16]) and Oskar Morgenstern published an *axiomatic* basis for utility theory.[17] What this means is the following: as long as the preferences exhibited by an individual satisfy certain basic axioms that any rational agent should satisfy, then *necessarily* the choices made by that individual can be described as maximizing the expected value of a utility function. In short, *a rational agent acts so as to maximize expected utility*.

It's hard to overstate the importance of this conclusion. In many ways, artificial intelligence has been mainly about working out the details of how to build rational machines.

Let's look in a bit more detail at the axioms that rational entities are expected to satisfy. Here's one, called *transitivity*: if you prefer A to B and you prefer B to C, then you prefer A to C. This seems pretty reasonable! (If you prefer sausage pizza to plain pizza, and you prefer plain pizza to pineapple pizza, then it seems reasonable to predict that

you will choose sausage pizza over pineapple pizza.) Here's another, called *monotonicity*: if you prefer prize A to prize B, and you have a choice of lotteries where A and B are the only two possible outcomes, you prefer the lottery with the highest probability of getting A rather than B. Again, pretty reasonable.

Preferences are not just about pizza and lotteries with monetary prizes. They can be about anything at all; in particular, they can be about entire future lives and the lives of others. When dealing with preferences involving sequences of events over time, there is an additional assumption that is often made, called *stationarity*: if two different futures A and B begin with the same event, and you prefer A to B, you still prefer A to B after the event has occurred. This sounds reasonable, but it has a surprisingly strong consequence: the utility of any sequence of events is the sum of rewards associated with each event (possibly discounted over time, by a sort of mental interest rate).[18] Although this "utility as a sum of rewards" assumption is widespread—going back at least to the eighteenth-century "hedonic calculus" of Jeremy Bentham, the founder of utilitarianism—the stationarity assumption on which it is based is not a necessary property of rational agents. Stationarity also rules out the possibility that one's preferences might change over time, whereas our experience indicates otherwise.

Despite the reasonableness of the axioms and the importance of the conclusions that follow from them, utility theory has been subjected to a continual barrage of objections since it first became widely known. Some despise it for supposedly reducing everything to money and selfishness. (The theory was derided as "American" by some French authors,[19] even though it has its roots in France.) In fact, it is perfectly rational to want to live a life of self-denial, wishing only to reduce the suffering of others. Altruism simply means placing substantial weight on the well-being of others in evaluating any given future.

Another set of objections has to do with the difficulty of obtaining the necessary probabilities and utility values and multiplying them

together to calculate expected utilities. These objections are simply confusing two different things: choosing the rational action and choosing it *by calculating expected utilities*. For example, if you try to poke your eyeball with your finger, your eyelid closes to protect your eye; this is rational, but no expected-utility calculations are involved. Or suppose you are riding a bicycle downhill with no brakes and have a choice between crashing into one concrete wall at ten miles per hour or another, farther down the hill, at twenty miles per hour; which would you prefer? If you chose ten miles per hour, congratulations! Did you calculate expected utilities? Probably not. But the choice of ten miles per hour is still rational. This follows from two basic assumptions: first, you prefer less severe injuries to more severe injuries, and second, for any given level of injuries, increasing the speed of collision increases the probability of exceeding that level. From these two assumptions it follows mathematically—without considering any numbers at all—that crashing at ten miles per hour has higher expected utility than crashing at twenty.[20] In summary, maximizing expected utility may not require calculating any expectations or any utilities. It's a purely *external* description of a rational entity.

Another critique of the theory of rationality lies in the identification of the locus of decision making. That is, what things count as agents? It might seem obvious that humans are agents, but what about families, tribes, corporations, cultures, and nation-states? If we examine social insects such as ants, does it make sense to consider a single ant as an intelligent agent, or does the intelligence really lie in the colony as a whole, with a kind of composite brain made up of multiple ant brains and bodies that are interconnected by pheromone signaling instead of electrical signaling? From an evolutionary point of view, this may be a more productive way of thinking about ants, since the ants in a given colony are typically closely related. As individuals, ants and other social insects seem to lack an instinct for self-preservation as distinct from colony preservation: they will always throw themselves into battle against invaders, even at suicidal odds. Yet sometimes

humans will do the same even to defend unrelated humans; it is as if the species benefits from the presence of some fraction of individuals who are willing to sacrifice themselves in battle, or to go off on wild, speculative voyages of exploration, or to nurture the offspring of others. In such cases, an analysis of rationality that focuses entirely on the individual is clearly missing something essential.

The other principal objections to utility theory are empirical—that is, they are based on experimental evidence suggesting that humans are irrational. We fail to conform to the axioms in systematic ways.[21] It is not my purpose here to defend utility theory as a formal model of human behavior. Indeed, humans cannot possibly behave rationally. Our preferences extend over the whole of our own future lives, the lives of our children and grandchildren, and the lives of others, living now or in the future. Yet we cannot even play the right moves on the chessboard, a tiny, simple place with well-defined rules and a very short horizon. This is not because our *preferences* are irrational but because of the *complexity* of the decision problem. A great deal of our cognitive structure is there to compensate for the mismatch between our small, slow brains and the incomprehensibly huge complexity of the decision problem that we face all the time.

So, while it would be quite unreasonable to base a theory of beneficial AI on an assumption that humans are rational, it's quite reasonable to suppose that an adult human has roughly consistent preferences over future lives. That is, *if you were somehow able to watch two movies, each describing in sufficient detail and breadth a future life you might lead, such that each constitutes a virtual experience, you could say which you prefer, or express indifference.*[22]

This claim is perhaps stronger than necessary, if our only goal is to make sure that sufficiently intelligent machines are not catastrophic for the human race. The very notion of *catastrophe* entails a definitely-not-preferred life. For catastrophe avoidance, then, we need claim only that adult humans can recognize a catastrophic future when it is spelled out in great detail. Of course, human preferences have a much

more fine-grained and, presumably, ascertainable structure than just "non-catastrophes are better than catastrophes."

A theory of beneficial AI can, in fact, accommodate inconsistency in human preferences, but the inconsistent part of your preferences can never be satisfied and there's nothing AI can do to help. Suppose, for example, that your preferences for pizza violate the axiom of transitivity:

ROBOT: Welcome home! Want some pineapple pizza?
YOU: No, you should know I prefer plain pizza to pineapple.
ROBOT: OK, one plain pizza coming up!
YOU: No thanks, I like sausage pizza better.
ROBOT: So sorry, one sausage pizza!
YOU: Actually, I prefer pineapple to sausage.
ROBOT: My mistake, pineapple it is!
YOU: I already said I like plain better than pineapple.

There is no pizza the robot can serve that will make you happy because there's always another pizza you would prefer to have. A robot can satisfy only the consistent part of your preferences—for example, let's say you prefer all three kinds of pizza to no pizza at all. In that case, a helpful robot could give you any one of the three pizzas, thereby satisfying your preference to avoid "no pizza" while leaving you to contemplate your annoyingly inconsistent pizza topping preferences at leisure.

Rationality for two

The basic idea that a rational agent acts so as to maximize expected utility is simple enough, even if actually doing it is impossibly complex. The theory applies, however, only in the case of a single agent acting alone. With more than one agent, the notion that it's possible—at least in principle—to assign probabilities to the different