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.

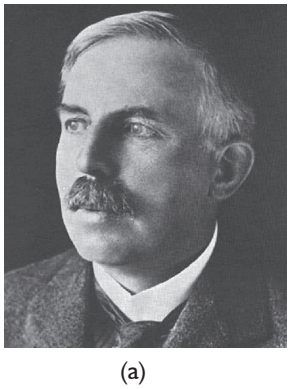
What Happens Next?

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.

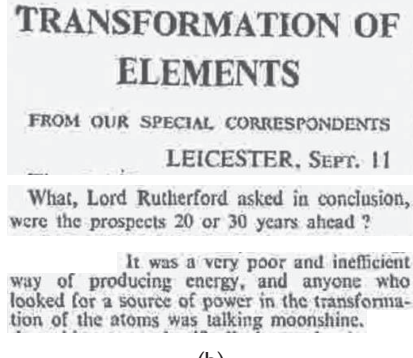
Scientific breakthroughs are notoriously hard to predict. To get a sense of just how hard, we can look back at the history of another field with civilization-ending potential: nuclear physics.

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.