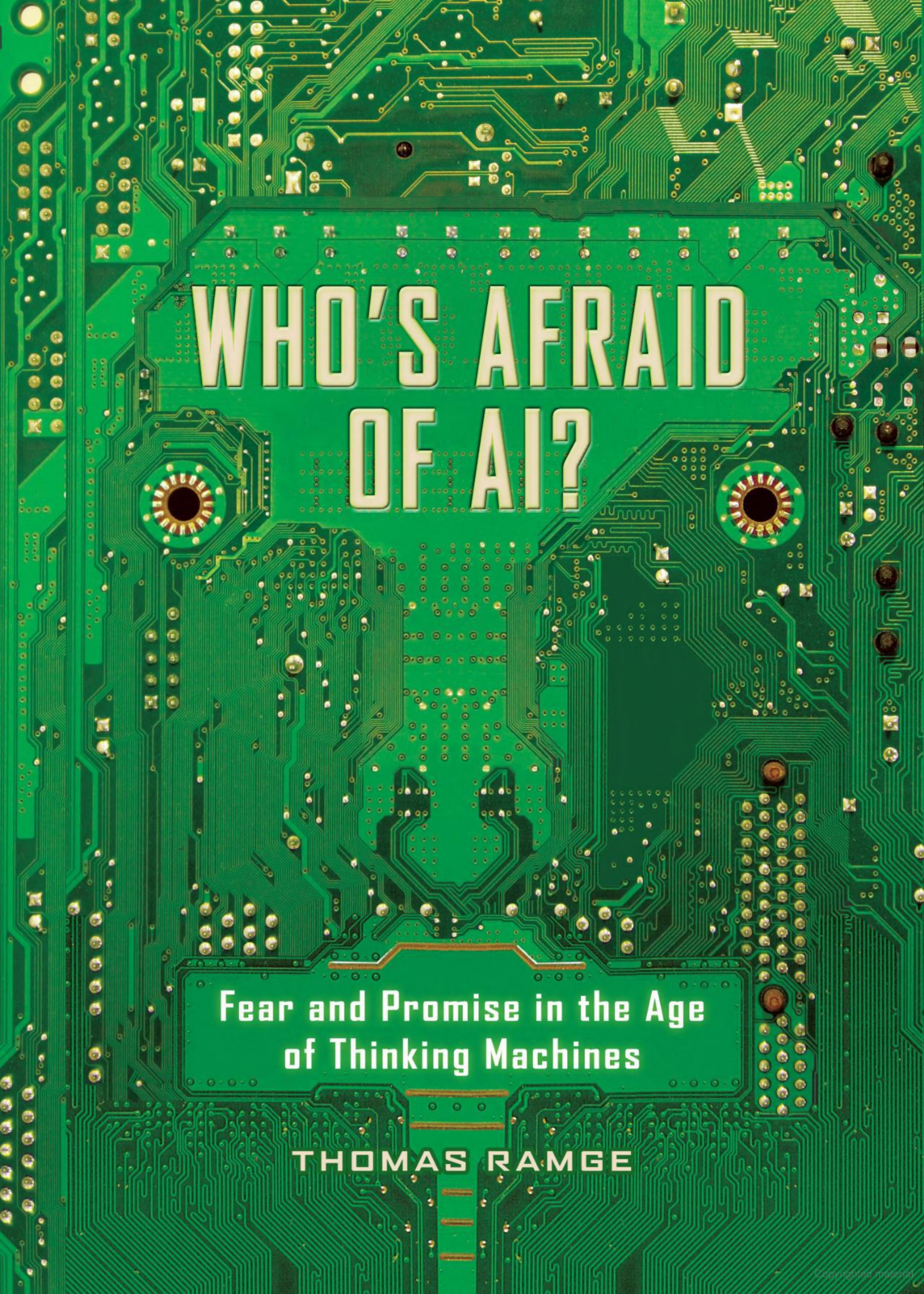


WHO'S AFRAID OF AI?

Fear and Promise in the Age of Thinking Machines

THOMAS RAMGE

WHO'S AFRAID OF AI?

Fear and Promise in the Age of Thinking Machines

THOMAS RAMGE

THE EXPERIMENT
NEW YORK

To my mother

Who's Afraid of AI?: *Fear and Promise in the Age of Thinking Machines*
Copyright © 2018 by Thomas Ramge
Translation copyright © 2019 by Thomas Ramge

Originally published in Germany as *Mensch und Maschine* by Philipp Reclam jun. Verlag GmbH & Co., KG in 2018. First published in North America by The Experiment, LLC, in 2019.

The Experiment, LLC, 220 East 23rd Street, Suite 600, New York, NY 10010-4658
theexperimentpublishing.com

Library of Congress Cataloging-in-Publication Data

Names: Ramge, Thomas, 1971- author.
Title: Who's afraid of AI? : fear and promise in the age of thinking machines / Thomas Ramge.
Other titles: Mensch und Maschine. English
Description: New York, NY : The Experiment, LLC, 2019. | "Originally published in Germany as Mensch und Maschine by Philipp Reclam jun. Verlag GmbH & Co., KG in 2018"--Title page verso. | Includes bibliographical references.
Identifiers: LCCN 2018055418 (print) | LCCN 2018056654 (ebook) | ISBN 9781615195510 (ebook) | ISBN 9781615195503 (pbk.)
Subjects: LCSH: Artificial intelligence--Popular works. | Machine learning--Popular works.
Classification: LCC Q315 (ebook) | LCC Q315 .R345 2019 (print) | DDC 006.3--dc23
LC record available at https://lccn.loc.gov/2018055418

ISBN 978-1-61519-550-3
Ebook ISBN 978-1-61519-551-0

Cover and text design by Beth Bugler | Spot drawings by Dinara Galieva
Translated by Jonathan Green

Manufactured in the United States of America
First printing April 2019
10 9 8 7 6 5 4 3 2 1

CONTENTS

INTRODUCTION 1
The Kitty Hawk Moment: Why Everything Is
About to Start Happening Fast

CHAPTER 1. 9
The Next Step of Automation: Machines Making Decisions

CHAPTER 2. 31
Turing's Heirs: A (Very) Short History of Artificial Intelligence

CHAPTER 3. 49
How Machines Learn to Learn: Artificial Neural Networks,
Machine Learning, and Feedback Effects

CHAPTER 4 63
Human Asks, Machine Answers: AI as a Daily Assistant,
Salesperson, Lawyer, and Doctor

CHAPTER 5 81
Robots as Coworkers: Smart Machines, Cobots,
and the Internet of Intelligent Things

CHAPTER 6 101
Superintelligence and the Singularity: Will Robots Seize Control?

Selected Sources 118
Acknowledgments 121
About the Author 122

> “I confess that in 1901,
> I said to my brother
> Orville that man would
> not fly for fifty years.”
>
> —Wilbur Wright

INTRODUCTION

The Kitty Hawk Moment: Why Everything Is About to Start Happening Fast

A million-dollar prize. A one-hundred-fifty-mile route through a restricted military area in the Mojave Desert. Those were the conditions for the US Department of Defense's first Defense Advanced Research Projects Agency (DARPA) Grand Challenge for autonomous vehicles in 2004. About a hundred teams entered. The best entry ground to a halt after seven and a half miles. Eight years later, in 2012, Google issued an inconspicuous press release: Its robot vehicles, famous from YouTube videos, had covered hundreds of thousands of accident-free miles in street traffic. Today, Tesla drivers have logged more than a billion miles on autopilot. To be sure, drivers have to take control of the steering wheel every now and then in tricky situations, the arising of which the autopilot duly makes them aware. But a seemingly unsolvable problem has in principle been solved. Despite several setbacks in the form of self-driving car accidents in 2018, the path to a fully automated vehicle for the masses is merely a matter of scale and fine-tuning.

Artificial intelligence is having its Kitty Hawk moment. For decades, the pioneers of aviation promised grandiose feats, only to fall short again and again. But then the Wright brothers had a breakthrough—their first flight in Kitty Hawk, North Carolina, in 1903—and the technology took off. Suddenly, what had for years been nothing but a boastful claim now worked.

And so it is for AI: After many years of relatively slow, underwhelming progress, the technology is finally starting to perform, and now a cascade of breakthroughs are flooding the market, with many more in the works. Computer programs' ability to recognize human faces has recently surpassed our own. Google Assistant can mimic a human voice and set a haircut appointment with such perfection that the person on the other end of the line has no idea they are talking to a data-rich IT system. In identifying certain cancer cells, computers today are already more accurate than the best doctors in the world—to say nothing of average doctors working in mediocre hospitals. Computers can now beat us at the near-infinitely complex board game Go, and if that weren't enough, they've also become better bluffers than the best poker players in the world. At the Japanese insurance company Fukoku Mutual, AI based on IBM's Watson system calculates reimbursements for medical bills according to each insurance contract's individual terms. At Bridgewater, the world's largest hedge fund, algorithms do much more than merely make decisions about investments. A system fed extensive employee data has become the robo-boss: It knows what is likely the best business strategy and the best team composition for particular tasks, and it makes recommendations for promotions and layoffs.

AI is the next step in automation. Heavy equipment has been doing our dirty work for a long time. Manufacturing robots have been getting more adept since the 1960s. Until now, however, IT systems have assisted with only the most routine knowledge work. But with artificial intelligence, machines are now making complex decisions that only human beings had been able to make. Or to state it more precisely: If the underlying data and the decision-making framework are sound, AI systems will make better decisions more quickly and less expensively than truck drivers, administrative staff members, sales clerks, doctors, investment bankers, and human resource managers, among others.

By twenty years after the first powered flight at Kitty Hawk, a new industry had arisen. Soon after that, air travel fundamentally changed the world. Artificial intelligence might follow a similar course. As soon as computer programs that learn from data prove themselves more efficient at a job than people are, their dominance of that industry will be inevitable. When built into physical machines like cars, robots, and drones, they take older automation processes in the material world to the next level. Networked together, they become an internet of intelligent things capable of cooperating with each other.

Gill Pratt, head of the Toyota Research Institute, makes a historical leap even farther than the dunes of Kitty Hawk in the Outer Banks. Pratt compares the most recent advances in AI to evolutionary biology's Cambrian explosion five hundred and forty million years ago. Almost all animal phyla originated during that period, setting off a kind of evolutionary arms race as the first complex species evolved the ability to see (among other things). With eyes, new habitats could be conquered and new biological niches could be exploited. Biodiversity exploded. The emphasis on vision is important: With biological breakthroughs in digital image recognition, AI now has eyes, too, allowing it to navigate—and learn from—its environment far more perceptively. MIT's Erik Brynjolfsson and Andrew McAfee continue the evolutionary comparison: "We expect to see a variety of new products, services, processes, and organizational forms and also numerous extinctions. There will certainly be some weird failures along with unexpected successes."

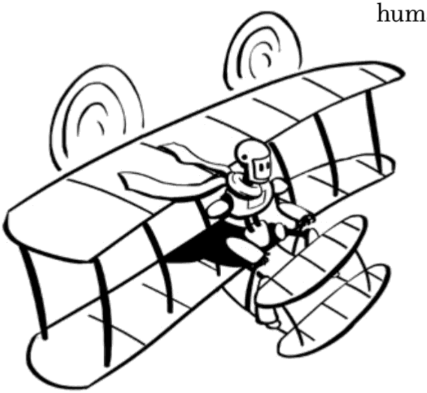
AI researchers and the producers of learning software systems have a powerful current pushing them forward right now. Startups in need of capital tend to paste the artificial intelligence label on every digital application, often without any consideration of whether the system actually learns from data and examples and can extrapolate from its experiences, or whether it is de facto traditionally programmed and mindlessly follows instructions. AI sells, and many buyers—whether research sponsors, investors, or users—are able to assess a product's technical operating principles only with difficulty. A magical aura currently surrounds AI—and not for the first time.

Artificial intelligence has already been through several hype cycles. Big promises have always been followed by phases of major disappointment. During these so-called AI winters, doubts have emerged even among AI's fervent disciples over whether they were chasing pipe dreams, driven by visions inspired by the science fiction authors whose books they had devoured as teenagers.

With all this in mind, we can still safely say that research on artificial intelligence has made breakthroughs on problems it had been dashing itself against for decades. And we'd probably give AI even more credit if we didn't so often take it for granted. When a machine multiplies better than a mathematical genius, plays chess more cunningly than the reigning world champion, or reliably shows us our way through a city, we are impressed for a short time. But as soon as calculators, chess programs, and navigation apps are inexpensive products for the masses, we perceive the technology as mundane. When AI comes into its own, we have a habit of seeing rote work where we once imagined feats of intelligence.

Today, the learning curve for machines appears to be sharply steeper than it is for human beings, which is fundamentally changing the relationship between humans and machines. Euphoric utopians in Silicon Valley like the author and Google researcher Ray Kurzweil see in this the key to solving all the major problems of our time, when a wish-granting artificial general intelligence (AGI) will make our lives easier, and maybe even eternal—in the form of an upload to the cloud, as some pundits believe. Apocalypticists, who—like the Oxford philosopher Nick Bostrom—are often European, fear the seizure of power by superintelligent machines and the end of humanity. Extreme positions make good headlines. For those who advocate them, extreme positions are good business in the market for our attention. Yet these positions are nevertheless important because they are leading many people to take a closer look.

Whoever wants to explore the opportunities and risks of a new technology first needs to understand the basics. They have to find comprehensible answers to these questions: What is artificial intelligence, anyway? What is it capable of today, and what will it be capable of in the foreseeable future? And what abilities will people need to develop if machines continue to become more and more intelligent? As we find more and more precise answers to these questions, we will get ready to address the big ones: Should we be afraid of AI? Should we fear humans using AI with malicious intent? And what kind of technological framework do humans have to set in place so that thinking machines—as agents of automation—can keep their promise to make the world wealthier and safer?



“Intelligence is what you use when you don’t know what to do.”

—Jean Piaget, biologist and
developmental psychologist

CHAPTER 1

The Next Step of Automation: Machines Making Decisions

Recognition, Insight, Action

The Tesla in autopilot is driving at eighty miles per hour in the highway's left lane. Ahead in the right lane, several trucks are driving at fifty-five miles per hour. The Tesla nears the column of trucks. The truck at the end of the convoy puts on its left blinker to signal that it wants to pass. The autopilot has to make a complex decision. Should the Tesla keep driving at the same speed or even accelerate in order to make sure it can pass the truck before the truck possibly changes lanes? Should it honk to warn the truck driver? Or should the Tesla, for safety's sake, brake and politely allow the truck to complete its passing maneuver at the cost of an increase in travel time? Braking would only be safe if there's not a lead-footed sports car driver tailgating six feet behind the Tesla, of course.

A few years ago, we would not have trusted this decision to a machine under any circumstances—and with complete justification. The technology had not yet proved that, statistically speaking, it was more likely to take us safely to our destination than we ourselves would if we were sitting behind the wheel using our own familiarity with traffic rules, knowledge based on experience, ability to anticipate human behavior, and famous gut instinct.

Today, Tesla drivers delegate many driving decisions to the computer. This is not without risks. Autonomous driving is far from perfect, whether at Tesla, Google, or the traditional car companies, which tirelessly work on autopilot systems but have not yet enabled many of their functions for safety reasons. In good weather and on clearly marked highways, today's machines are demonstrably the better drivers. It is only a question of time until this is also true in the city or at night or in fog, or until a machine decides not to drive in black-ice conditions at all because the risks are simply too high.

As the old saying in AI research goes, what's hard for people is easy for machines, and vice versa. Driving a car, which involves thousands of small but nevertheless complex decision scenarios during each trip, was previously impossible for computers. Why is that changing now? In abstract terms, it is because software that learns from data in connection with controllable hardware has increasingly mastered three core skills—recognition, insight, and implementation of an action.

In the example of the Tesla and the truck with a blinking turn signal, this means that GPS navigation, high-resolution cameras, and laser and radar sensors inform the system of even more than exactly where the car is, how fast the truck is going, the condition of the road, and if there is an emergency lane to the right. The system's image recognition software can also reliably identify that it is the truck's turn signal that is blinking and not a lamp at a construction site somewhere in the distance. Computers have gained this ability to recognize things only in the last few years. The best of them today can distinguish between crumpled paper that the vehicle can safely drive over and a rock that needs to be driven around.

All visual (and other sensory) data flow into a small supercomputer, the car's artificial brain, composed of many computing processors (called cores) and graphics processors. The processing unit has to sort the information in fractions of a second as it simultaneously synchronizes real-time data with previously collected data and rules that have been programmed into the system. The Tesla system knows that it has the right-of-way in this instance. It was equipped with the traffic rule that the truck driver is only allowed to change lanes and pass if there is no one approaching from behind. Fortified by machine learning from many billions of miles in street traffic—the so-called feedback data—the system also knows that truck drivers do not always follow traffic rules. There is a significant probability that the truck will switch lanes even though the Tesla is approaching from behind, and the car knows as well that it is not at all in the best interests of its passengers if a robot car insists upon following traffic rules when it risks a serious accident by doing so.

From the observed circumstances, programmed rules, and prior experience, the system deduces the best possibility among the many computable scenarios for avoiding an accident while still moving forward quickly. At heart, it is a cognitive decision, the choice of one course of action among many. The best solution to the problem is a probability calculation that draws on many variables.

A partially automatic assisted driving system offers its insights to the driver only as a basis for decisions, by sounding a warning beep, for example, if a truck not only signals but also makes small swerving motions indicating that the driver is truly about to turn the steering wheel to the left. The human driver can then follow the machine-derived advice or ignore it. But an autopilot worthy of the name turns its insights directly into action. It brakes or honks or drives on stoically. The computer is able to implement its decision because an autonomous vehicle is a highly developed cyber-physical system. The digital system controls the functions of the physical machine, such as the gas, brakes, and steering, with great skill. An airplane's autopilot can take off or land in normal conditions more precisely than any pilot with a captain's cap on his or her head. With a completely digital system, such as a trading bot, used for high-frequency stock market trading, implementing the decision, naturally, takes place purely digitally, but the automation principle is the same: recognize patterns in the data, deduce insights from statistics and algorithms, and implement an insight as a decision through a technological response. The machine scouts the market for trends, sees an opportunity for an advantageous trade, and clicks on "Buy it now."

Polanyi's Paradox

Measuring the effects of decisions and including the results in future decision-making is arguably the essence of artificial intelligence systems. They make decisions on the basis of feedback loops. If the Tesla in the situation described here causes an accident, it transmits this feedback back to the central computer, and all other Teslas will (hopefully) drive more defensively in comparable situations. If AI software for approving loans registers too many defaults, it will tighten the criteria for subsequent loan applicants. If a harvester receives feedback that it is picking too many unripe apples, on the next pass it will be able to make better decisions as to what ratio of red to green on an apple's surface is sufficient. The essential difference between artificial intelligence and classic IT systems lies in this ability of AI to independently improve its own calculations by classifying the results of its actions. Autocorrection is built into the system.

Since the first mainframe computers of the 1940s, programming a computer has meant that a human being painstakingly teaches a theoretical model to a machine. The model contains particular rules that the machine can apply. If the machine is fed data that fits particular tasks or questions, then it can usually solve them more quickly, precisely, and reliably than a human being can. In essence, classical programming involves transferring existing knowledge from programmers' heads into a machine. This technological approach has a natural limit: A large part of our knowledge is implicit.

It is true that we can recognize faces, but we don't know exactly how we do it. Evolution has given us this ability, but we don't have a good theory for why we are immediately able to identify Beyoncé or George Clooney, even if the light is bad and the face is half covered. It's also almost impossible to exactly describe the best way to teach a child to ski or swim. Another famous example of implicit knowledge is the answer to the question, *What is hard-core pornography?* Supreme Court Justice Potter Stewart, struggling for a legally watertight definition, found only the despairing answer "I know it when I see it." This problem has a name: Polanyi's paradox. It describes a limit that until now had seemed insurmountable for software programmers. Without theory, broken down into rules, we can't impart our knowledge and our abilities to machines.

Artificial intelligence overcomes Polanyi's paradox by having human beings create only the framework in which a machine learns how to learn. There are countless competing methods and approaches among the various schools of AI. The majority of them, however—including the most important and successful ones—follow the basic principle of giving computers not so much theories or rules, but rather goals. Computers learn how to reach these goals in a training phase involving many examples and feedback as to whether or not they have attained the goals set by human beings.

This raises the question of whether machine learning via feedback loops should be considered a form of intelligence. Many AI researchers don't especially like the concept of "artificial intelligence," preferring instead to use the designation *machine learning*.

Strong and Weak AI

The term *artificial intelligence* has been controversial since the computer pioneers associated with Marvin Minsky coined it in 1956 at their famous Dartmouth conference (more on this later). And scientists still don't even agree on what constitutes *human* intelligence. Can such a concept be appropriate for machines at all? Discussions of artificial intelligence quickly drift into very fundamental questions. For example: *Is thought without consciousness possible?* Or, *Will machines soon be more intelligent than people, and will they develop the ability to make themselves more and more intelligent, possibly developing a self-image and consciousness and their own interests in the process?* If so, will we have to grant human rights to thinking machines? Or will humans and machines just converge and form transhumanist beings escalating humanity to the next stage of evolution?

These are questions about so-called strong AI (often called general AI)—cognitively advanced, humanlike AI—and they're important. The long-term consequences of such technology should be carefully considered while it develops, not in retrospect, as in the case of, say, nuclear weapons. The last chapter of this book will touch on these questions. But these concerns are far out on the horizon. Much more urgent is the matter of weak AI, AI that is technologically possible today and in the foreseeable future. But first, let's clarify exactly what we mean by weak AI (narrow AI).

The American philosopher of language John R. Searle suggested distinguishing between strong and weak AI about four decades ago. For the time being, strong AI is science fiction. Weak artificial intelligence, on the other hand, is at work in the here and now whenever a computer system completes a task that until recently we thought only a human being exerting his or her brain in some fashion could manage—for example, case work for insurance companies, or the writing of news or sports stories.

Embedded in physical machines, AI enhances the intelligence of not only automobiles, but also factories, farm equipment, drones, and rescue and caregiving robots. But about that word *intelligence*: We can describe AI in the language we use to describe ourselves, but it's important to keep in mind that to complete a task, smart machines don't have to imitate human approaches—or the biochemical processes in the human brain in any sense at all. They typically have the ability to search autonomously for mathematical solutions, to improve the algorithms they are given, and order to suggest a likely successful therapy for an illness or injury, compared to highly qualified doctors needing one hundred sixty hours to perform the same analysis, then the return is measured not in dollars, but in human lives saved.

"Artificial intelligence will change the world like electricity did." This sentiment, or something close to it, appears in many articles and studies concerning AI. In times of technological paradigm shifts, experts' predictions—especially ones tending toward euphoria—should be treated with caution. The future can only halfway reliably be predicted from the data of the past if nothing fundamental changes.

In this respect, digitalization itself creates an interesting paradox. More data and analysis raise people's ability to forecast the future. But the disruptive nature of digital technology creates unpredictable change. And yet we are on solid ground with the hypothesis that intelligent machines will fundamentally shake up our life, our work, our economy, and our society in the coming two decades. The analogy to the introduction of electricity is correct insofar as systems that learn from data represent an integrative technology. Like the combustion engine, the development of plastics, or the Internet, it has an effect in many areas and simultaneously creates the requisite conditions for new innovations whose appearance and effects we can't even imagine today.

Electricity made possible efficient trains, the assembly line, light for libraries, the telephone, the film industry, the microwave, computers, and the battery-driven explorations of a Mars rover over rugged extraterrestrial terrain. We can't imagine modern life without electricity. Andrew Ng, a Stanford University professor and former head of the AI teams at Google and Baidu, tackles the question of which sectors AI will affect: "It might be easier to think about what industries AI will *not* transform." This is no longer a statement about the future. It describes the present, including both positive aspects and disconcerting ones.

Rage Against the Machine?

No one can reliably predict today whether artificial intelligence systems will primarily destroy human jobs or if their second wave will create new work, as has been the case in earlier technological revolutions. The machine-wrecking Luddites of the early nineteenth century smashed the first mechanical looms in central England with sledgehammers. Rage against the machine! Destroy whatever destroys you. But their rage was of little use to them. While productivity and the gross domestic product rose rapidly, for them, working conditions deteriorated. It took decades before the return on investment in automation reached their children and tasks or how dynamically they will spread. The quandary regarding making a solid prognosis is at heart a question of speed. The faster AI extends itself into the human workplace, the less time remains for people to adjust their individual qualifications and their collective safety systems. A new generation of people who lose out to automation then becomes more likely. Even with all the uncertainty in the forecast, however, it is certain that politicians worldwide have until now found only a few intelligent answers to the challenges of the next major wave of automation. We are not well prepared for the return of the machinery question.

The Flaw in the Machine

The even more pressing question for humanity may well be: Will a superintelligence emerge that autonomously and in feedback loops calculates an increasingly better understanding of the world and of itself? An AI system that leads to humanity being "deposed from its position as apex cogitator," as Nick Bostrom, head of the Future of Humanity Institute at Oxford, puts it. The consequence would be that humans could no longer control the superintelligent system. And might this superintelligence even turn against humanity like in science fiction, so that in the end the machine exterminates human beings?