REBOOTING AI

Building Artificial Intelligence We Can Trust

GARY MARCUS
and ERNEST DAVIS

REBOOTING AI

BUILDING ARTIFICIAL INTELLIGENCE WE CAN TRUST

Gary Marcus and Ernest Davis

Pantheon Books New York

Grateful acknowledgment is made to Houghton Mifflin Harcourt Publishing Company for permission to reprint an excerpt from "A Little Girl Tugs at the Tablecloth," from *Monologue of a Dog: New Poems by Wislawa Szymborska*, translated from the Polish by Stanislaw Baranczak and Clare Cavanagh.

Library of Congress Cataloging-in-Publication Data
Names: Marcus, Gary, author. Davis, Ernest, author.
Title: Rebooting AI : building artificial intelligence we can trust /
Gary Marcus and Ernest Davis.
Description: First edition. New York : Pantheon Books, 2019.
Includes bibliographical references and index.
Identifiers: LCCN 2019005842. ISBN 9781524748258 (hardcover : alk. paper).
ISBN 9781524748265 (ebook).
Subjects: LCSH: Artificial intelligence.
Classification: LCC Q335 .M368 2019 | DDC 006.3--dc23 |
LC record available at lccn.loc.gov/2019005842

www.pantheonbooks.com

Jacket image by Nadia Snopek/Shutterstock
Jacket design by Kelly Blair

Printed in the United States of America
First Edition

2 4 6 8 9 7 5 3 1

Contents

REBOOTING AI

1

Mind the Gap

Machines will be capable, within twenty years, of doing any work a man can do.

—AI PIONEER HERB SIMON, 1965

FIRST CHILD [*on a long, arduous journey*]: Is it much further, Papa Smurf?

FATHER: Not far now.

SECOND CHILD [*much later*]: Is it much further, Papa Smurf?

FATHER: Not far now.

—THE SMURFS

Since its earliest days, artificial intelligence has been long on promise, short on delivery. In the 1950s and 1960s, pioneers like Marvin Minsky, John McCarthy, and Herb Simon genuinely believed that AI could be solved before the end of the twentieth century. "Within a generation," Marvin Minsky famously wrote, in 1967, "the problem of artificial intelligence will be substantially solved." Fifty years later, those promises still haven't been fulfilled, but they have never stopped coming. In 2002, the futurist Ray Kurzweil made a public bet that AI would "surpass native human intelligence" by 2029. In November 2018 Ilya Sutskever, co-founder of OpenAI, a major AI research institute, suggested that "near term AGI [artificial general intelligence] should be taken seriously as a possibility." Although it is still theoretically possible that Kurzweil and Sutskever might turn out to be right, the odds against this happening are very long. Getting to that level—general-purpose artificial intelligence with the flexibility of human intelligence—isn't some small step from where we are now;

instead it will require an immense amount of foundational progress—not just more of the same sort of thing that's been accomplished in the last few years, but, as we will show, something entirely different.

Even if not everyone is as bullish as Kurzweil and Sutskever, ambitious promises still remain common, for everything from medicine to driverless cars. More often than not, what is promised doesn't materialize. In 2012, for example, we heard a lot about how we would be seeing "autonomous cars [in] the near future." In 2016, IBM claimed that Watson, the AI system that won at *Jeopardy!*, would "revolutionize healthcare," stating that Watson Health's "cognitive systems [could] understand, reason, learn, and interact" and that "with [recent advances in] cognitive computing . . . we can achieve more than we ever thought possible." IBM aimed to address problems ranging from pharmacology to radiology to cancer diagnosis and treatment, using Watson to read the medical literature and make recommendations that human doctors would miss. At the same time, Geoffrey Hinton, one of AI's most prominent researchers, said that "it is quite obvious we should stop training radiologists."

In 2015 Facebook launched its ambitious and widely covered project known simply as M, a chatbot that was supposed to be able to cater to your every need, from making dinner reservations to planning your next vacation.

As yet, none of this has come to pass. Autonomous vehicles may someday be safe and ubiquitous, and chatbots that can cater to every need may someday become commonplace; so too might superintelligent robotic doctors. But for now, all this remains fantasy, not fact.

The driverless cars that do exist are still primarily restricted to highway situations with human drivers required as a safety backup, because the software is too unreliable. In 2017, John Krafcik, CEO at Waymo, a Google spinoff that has been working on driverless cars for nearly a decade, boasted that Waymo would shortly have driverless cars with no safety drivers. It didn't happen. A year later, as *Wired* put it, the bravado was gone, but the safety drivers weren't. Nobody really thinks that driverless cars are ready to drive fully on their own in cities or in bad weather, and early optimism has been

replaced by widespread recognition that we are at least a decade away from that point—and quite possibly more.

IBM Watson's transition to health care similarly has lost steam. In 2017, MD Anderson Cancer Center shelved its oncology collaboration with IBM. More recently it was reported that some of Watson's recommendations were "unsafe and incorrect." A 2016 project to use Watson for the diagnosis of rare diseases at the Marburg, Germany, Center for Rare and Undiagnosed Diseases was shelved less than two years later, because "the performance was unacceptable." In one case, for instance, when told that a patient was suffering from chest pain, the system missed diagnoses that would have been obvious even to a first year medical student, such as heart attack, angina, and torn aorta. Not long after Watson's troubles started to become clear, Facebook's M was quietly canceled, just three years after it was announced.

Despite this history of missed milestones, the rhetoric about AI remains almost messianic. Eric Schmidt, the former CEO of Google, has proclaimed that AI would solve climate change, poverty, war, and cancer. XPRIZE founder Peter Diamandis made similar claims in his book *Abundance*, arguing that strong AI (when it comes) is "definitely going to rocket us up the Abundance pyramid." In early 2018, Google CEO Sundar Pichai claimed that "AI is one of the most important things humanity is working on . . . more profound than . . . electricity or fire." (Less than a year later, Google was forced to admit in a note to investors that products and services "that incorporate or utilize artificial intelligence and machine learning, can raise new or exacerbate existing ethical, technological, legal, and other challenges.")

Others agonize about the potential dangers of AI, often in ways that show a similar disconnect from current reality. One recent nonfiction bestseller by the Oxford philosopher Nick Bostrom grappled with the prospect of superintelligence taking over the world, as if that were a serious threat in the foreseeable future. In the pages of *The Atlantic*, Henry Kissinger speculated that the risk of AI might be so profound that "human history might go the way of

the Incas, faced with a Spanish culture incomprehensible and even awe-inspiring to them." Elon Musk has warned that working on AI is "summoning the demon" and a danger "worse than nukes," and the late Stephen Hawking warned that AI could be "the worst event in the history of our civilization."

But what AI, exactly, are they talking about? Back in the real world, current-day robots struggle to turn doorknobs, and Teslas driven in "Autopilot" mode keep rear-ending parked emergency vehicles (at least four times in 2018 alone). It's as if people in the fourteenth century were worrying about traffic accidents, when good hygiene might have been a whole lot more helpful.

One reason that people often overestimate what AI can actually do is that media reports often overstate AI's abilities, as if every modest advance represents a paradigm shift.

Consider this pair of headlines describing an alleged breakthrough in machine reading.

> "Robots Can Now Read Better than Humans, Putting Millions of Jobs at Risk"
>
> —*NEWSWEEK*, JANUARY 15, 2018

> "Computers Are Getting Better than Humans at Reading"
>
> —*CNN MONEY*, JANUARY 16, 2018

The first is a more egregious exaggeration than the second, but both wildly oversell minor progress. To begin with, there were no actual robots involved, and the test only measured one tiny aspect of reading. It was far from a thorough test of comprehension. No actual jobs were remotely in jeopardy.

All that happened was this. Two companies, Microsoft and Alibaba, had just built programs that made slight incremental progress on a particular test of a single narrow aspect of reading (82.65 percent versus the previous record of 82.136 percent), known as SQuAD

(the Stanford Question Answering Dataset), arguably achieving human-level performance on that specific task where they weren't quite at human level before. One of the companies put out a press release that made this minor achievement sound much more revolutionary than it really was, announcing the creation of "AI that can read a document and answer questions about it as well as a person."

Reality was much less sexy. Computers were shown short passages of text drawn from an exam designed for research purposes and asked questions about them. The catch is that in every case the correct answers appeared *directly in the text*—which rendered the exam an exercise in underlining, and nothing more. Untouched was much of the real challenge of reading: inferring meanings that are implied yet not always fully explicit.

Suppose, for example, that we hand you a piece of paper with this short passage:

> Two children, Chloe and Alexander, went for a walk. They both saw a dog and a tree. Alexander also saw a cat and pointed it out to Chloe. She went to pet the cat.

It is trivial to answer questions like "Who went for a walk?," in which the answer ("Chloe and Alexander") is directly spelled out in the text, but any competent reader should just as easily be able to answer questions that are not directly spelled out, like "Did Chloe see the cat?" and "Were the children frightened by the cat?" If you can't do that, you aren't really following the story. Because SQuAD didn't include any questions of this sort, it wasn't really a strong test of reading; as it turns out the new AI systems would not have been able to cope with them.* By way of contrast, Gary tested the story

* Even easier questions like "What did Alexander see?" would be out of bounds, because the answer (a dog, a tree, and a cat) requires highlighting two pieces of text that aren't contiguous, and the test had made it easy on machines by restricting questions to those that could be answered with a single bit of contiguous text.

on his daughter Chloe, then four and a half years old, and she had no trouble making the inference that the fictitious Chloe had seen a cat. (Her older brother, then not quite six years old, went a step further, musing about what would happen if the dog actually turned out to be a cat; no current AI could begin to do that.)

Practically every time one of the tech titans puts out a press release, we get a reprise of this same phenomenon, in which a minor bit of progress is portrayed in many (mercifully not all) media outlets as a revolution. A couple of years ago, for example, Facebook introduced a bare-bones proof-of-concept program that read simple stories and answered questions about them. A slew of enthusiastic headlines followed, like "Facebook Thinks It Has Found the Secret to Making Bots Less Dumb" (*Slate*) and "Facebook AI Software Learns and Answers Questions. Software able to read a synopsis of Lord of the Rings and answer questions about it could beef up Facebook search" (*Technology Review*).

That really would be a major breakthrough—if it were true. A program that could assimilate even the *Reader's Digest* or Cliffs-Notes versions of Tolkien (let alone the real thing) would be a major advance.

Alas, a program genuinely capable of such a feat is nowhere in sight. The synopsis that the Facebook system actually read was just four lines long:

> Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring. Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring. Frodo journeyed to Mount Doom. Frodo dropped the ring there. Sauron died. Frodo went back to the Shire. Bilbo travelled to the Grey Havens. The End.

And even then, all the program could do was answer basic questions directly addressed in those sentences, such as "Where is the ring?," "Where is Bilbo now?," and "Where is Frodo now?" Forget about asking why Frodo dropped the ring.

The net effect of a tendency of many in the media to overreport

technology results is that the public has come to believe that AI is much closer to being solved than it really is.

Whenever you hear about a supposed success in AI, here's a list of six questions you could ask:

1. Stripping away the rhetoric, what did the AI system actually do here?
2. How general is the result? (E.g., does an alleged reading task measure all aspects of reading, or just a tiny slice of it?)
3. Is there a demo where I can try out my own examples? (Be very skeptical if there isn't.)
4. If the researchers (or their press people) allege that an AI system is better than humans, then which humans, and how much better?
5. How far does succeeding at the particular task reported in the new research actually take us toward building genuine AI?
6. How robust is the system? Could it work just as well with other data sets, without massive amounts of retraining? (E.g., could a game-playing machine that mastered chess also play an action-adventure game like Zelda? Could a system for recognizing animals correctly identify a creature it had never seen before as an animal? Would a driverless car system that was trained during the day be able to drive at night, or in the snow, or if there was a detour sign not listed on its map?)

This book is about how to be skeptical, but more than that, it's about why AI, so far, hasn't been on the right track, and what we might do to work toward AI that is robust and reliable, capable of functioning in a complex and ever-changing world, such that we can genuinely trust it with our homes, our parents and children, our medical decisions, and ultimately our lives.

To be sure, AI has been getting more impressive, virtually every day, for the last several years, sometimes in ways that are truly amazing.

There have been major advances in everything from game playing to speech recognition to identifying faces. A startup company we are fond of, Zipline, uses a bit of AI to guide drones to deliver blood to patients in Africa, a fantastic application that would have been out of the question just a few years ago.

Much of this recent success in AI has been driven largely by two factors: first, advances in hardware, which allow for more memory and faster computation, often by exploiting many machines working in parallel; second, big data, huge data sets containing gigabytes or terabytes (or more) of data that didn't exist until a few years ago, such as ImageNet, a library of 15 million labeled pictures that has played a pivotal role in training computer vision systems; Wikipedia; and even the vast collections of documents that together make up the World Wide Web.

Emerging in tandem with the data has been an algorithm for churning through that data, called *deep learning*, a kind of powerful statistical engine that we will explain and evaluate in chapter 3. Deep learning has been at the center of practically every advance in AI in the last several years, from DeepMind's superhuman Go and chess player AlphaZero to Google's recent tool for conversation and speech synthesis, Google Duplex. In each case, big data plus deep learning plus faster hardware has been a winning formula.

Deep learning has also been used with substantial success for a wide range of practical applications from diagnosing skin cancer to predicting earthquake aftershocks to detecting credit card fraud. It's also been used in art and music, as well as for a huge number of commercial applications, from deciphering speech to labeling photos to organizing people's news feeds. You can use deep learning to identify plants or to automatically enhance the sky in your photos and even to colorize old black-and-white pictures.

Along with deep learning's stunning success, AI has become a huge business. Companies like Google and Facebook are in an epic battle for talent, often paying PhDs the sort of starting salaries we expect for professional athletes. In 2018, the most important

scientific conference for deep learning sold out in twelve minutes. Although we will be arguing that AI with human-level flexibility is much harder than many people think, there is no denying that real progress has been made. It's not an accident that the broader public has become excited about AI.

Nations have, too. Countries like France, Russia, Canada, and China have all made massive commitments to AI. China alone is planning to invest $150 billion by 2030. The McKinsey Global Institute estimates that the overall economic impact of AI could be $13 trillion, comparable to the steam engine in the nineteenth century and information technology in the twenty-first.

Still, that doesn't guarantee that we are on the right path.

Indeed, even as data has become more plentiful, and clusters of computers have become faster, and investments bigger, it is important to realize that something fundamental is still missing. Even with all the progress, in many ways machines are still no match for people.

Take reading. When you read or hear a new sentence, your brain, in less than a second, performs two types of analysis: (1) it parses the sentence, deconstructing it into its constituent nouns and verbs and what they mean, individually and collectively; and (2) it connects that sentence to what you know about the world, integrating the grammatical nuts and bolts with a whole universe of entities and ideas. If the sentence is a line of dialogue in a movie, you update your understanding of a character's intentions and prospects. Why did they say what they said? What does that tell us about their character? What are they trying to achieve? Is it truthful or deceptive? How does it relate to what has happened before? How will their speech affect others? For example, when thousands of former slaves stand up one by one and declare "I am Spartacus," each one risking execution, we all know instantly that every one of them (except Spartacus himself) is lying, and that we have just witnessed something moving and profound. As we will demonstrate, current AI programs can't

do or understand anything remotely like this; as far as we can tell, they aren't even on track to do so. Most of the progress that has been made has been on problems like object recognition that are entirely different from challenges in understanding meaning.

The difference between the two—object recognition and genuine comprehension—matters in the real world. The AI programs that power the social media platforms we have now, for example, can help spread fake news, by feeding us outrageous stories that garner clicks, but they can't understand the news well enough to judge which stories are fake and which are real.

Even the prosaic act of driving is more complex than most people realize. When you drive a car, 95 percent of what you do is absolutely routine and easily replicated by machines, but the first time a teenager darts out in front of your car on a battery-powered hoverboard, you will have to do something no current machine can do reliably: reason and act on something new and unexpected, based not on some immense database of prior experience, but on a powerful and flexible understanding of the world. (And you can't just slam on the brakes every time you see something unexpected, or you could get rear-ended every time you stop for a pile of leaves in the road.)

Currently, truly driverless cars can't be counted on. Perhaps the closest thing commercially available to consumers is Autopilot-equipped Teslas, but they still demand the full attention of the human driver at all times. The system is reasonably reliable on highways in good weather, but less likely to be reliable in dense urban areas. On a rainy day in the streets of Manhattan or Mumbai, we would still trust our lives sooner to a randomly chosen human driver than to a driverless car.* The technology just isn't mature yet. As a

* Directly comparable data for comparing the safety of humans versus machines have not yet been published. Much of the testing has been done on highways, which are easiest for machines, rather than in crowded urban areas, which pose greater challenges for AI. Published data suggest that the most reliable extant software requires human intervention roughly once every 10,000 miles, even in rather easy driving conditions. By way of imperfect comparison, human drivers are involved in fatal accidents on average

Toyota vice president for automated driving research recently put it, "Taking me from Cambridge to Logan Airport with no driver in any Boston weather or traffic condition—that might not be in my lifetime."

Likewise, when it comes to understanding the plot of a movie or the point of a newspaper article, we would trust middle school students over any AI system. And, much as we hate changing diapers, we can't imagine any robot now in development being reliable enough to help.

The central problem, in a word: current AI is *narrow;* it works for particular tasks that it is programmed for, provided that what it encounters isn't too different from what it has experienced before. That's fine for a board game like Go—the rules haven't changed in 2,500 years—but less promising in most real-world situations. Taking AI to the next level will require us to invent machines with substantially more flexibility.

What we have for now are basically digital idiots savants: software that can, for example, read bank checks or tag photos or play board games at world champion levels, but does little else. Riffing off a line from investor Peter Thiel about wanting flying cars and instead getting 140 characters, we wanted Rosie the Robot, ready at a moment's notice to change our kids' diapers and whip up dinner, and instead we got Roomba, a hockey-puck-shaped vacuum cleaner with wheels.

Or consider Google Duplex, a system that makes phone calls and sounds remarkably human. When it was announced in spring 2018 there was plenty of discussion about whether computers should be required to identify themselves when making such phone calls.

---

only once in every 100 million miles. One of the greatest risks in driverless cars is that if the machine requires intervention only infrequently, we are apt to tune out, and not be available quickly enough when intervention is required.

(Under much public pressure, Google agreed to that after a couple of days.) But the real story is how narrow Duplex was. For all the fantastic resources of Google (and its parent company, Alphabet), the system that they created was so narrow it could handle just three things: restaurant reservations, hair salon appointments, and the opening hours of a few selected businesses. By the time the demo was publicly released, on Android phones, even the hair salon appointments and the opening hour queries were gone. Some of the world's best minds in AI, using some of the biggest clusters of computers in the world, had produced a special-purpose gadget for making nothing but restaurant reservations. It doesn't get narrower than that.

To be sure, that sort of narrow AI is certainly getting better by leaps and bounds, and undoubtedly there will be more breakthroughs in the years to come. But it's also telling: AI could and should be about so much more than getting your digital assistant to book a restaurant reservation.

It could and should be about curing cancer, figuring out the brain, inventing new materials that allow us to improve agriculture and transportation, and coming up with new ways to address climate change. At DeepMind, now part of Alphabet, there used to be a motto, "Solve intelligence, and then use intelligence to solve everything else." While we think that might have been overpromising a bit—problems are often political rather than purely technical—we agree with the sentiment; progress in AI, if it's large enough, can have major impact. If AI could read and reason as well as humans—yet work with the precision and patience and massive computational resources of modern computer systems—science and technology might accelerate rapidly, with huge implications for medicine and the environment and more. That's what AI should be about. But, as we will show you, we can't get there with narrow AI alone.

Robots, too, could have a much more profound impact than they currently do, if they were powered by a deeper kind of AI than we currently have. Imagine a world in which all-purpose domestic robots have finally arrived, and there are no longer windows to wash,

floors to sweep, and, for parents, lunches to pack and diapers to clean. Blind people could use robots as assistants; the elderly could use them as caretakers. Robots could also take over jobs that are dangerous or entirely inaccessible for people, working underground, underwater, in fires, in collapsed buildings, in mine fields, or in malfunctioning nuclear reactors. Workplace fatalities could be greatly reduced, and our capacity to extract precious natural resources might be greatly improved, without putting humans at risk.

Driverless cars, too, could have a profound impact—if we could make them work reliably. Thirty thousand people a year die in the United States in auto accidents, and a million around the globe, and if the AI for guiding autonomous vehicles can be perfected those numbers could be greatly reduced.

The trouble is that the approaches we have now won't take us there, not to domestic robots, or automated scientific discoveries; they probably can't even take us to fully reliable driverless cars. Something important is still missing. Narrow AI alone is not enough.

Yet we are ceding more and more authority to machines that are unreliable and, worse, lack any comprehension of human values. The bitter truth is that for now the vast majority of dollars invested in AI are going toward solutions that are brittle, cryptic, and too unreliable to be used in high-stakes problems.

The core problem is trust. The narrow AI systems we have now often work—on what they are programmed for—but they can't be trusted with anything that hasn't been precisely anticipated by their programmers. That is particularly important when the stakes are high. If a narrow AI system offers you the wrong advertisement on Facebook, nobody is going to die. But if an AI system drives your car into an unusual-looking vehicle that isn't in its database, or misdiagnoses a cancer patient, it could be serious, even fatal.

What's missing from AI today—and likely to stay missing, until and unless the field takes a fresh approach—is *broad* (or "general")

It's no exaggeration to say that our future depends on it. AI has enormous potential to help us with some of the largest challenges that face humanity, in key areas including medicine, the environment, and natural resources. But the more power we hand off to AI, the more it becomes critical that AI use that power in ways that we can count on. And that means rethinking the whole paradigm.

We call this book *Rebooting AI* because we believe that the current approach isn't on a path to get us to AI that is safe, smart, or reliable. A short-term obsession with narrow AI and the easily achievable "low-hanging fruit" of big data has distracted too much attention away from a longer-term and much more challenging problem that AI needs to solve if it is to progress: the problem of how to endow machines with a deeper understanding of the world. Without that deeper understanding, we will never get to truly trustworthy AI. In the technical lingo, we may be stuck at a local maximum, an approach that is better than anything similar that's been tried, but nowhere good enough to get us where we want to go.

For now, there is an enormous gap—we call it "the AI Chasm"—between ambition and reality.

That chasm has roots in three separate challenges, each of which needs to be faced honestly.

The first we call the *gullibility gap*, which starts with the fact that we humans did not evolve to distinguish between humans and machines—which leaves us easily fooled. We attribute intelligence to computers because we have evolved and lived among human beings who themselves base their actions on abstractions like ideas, beliefs, and desires. The behavior of machines is often superficially similar to the behavior of humans, so we are quick to attribute to machines the same sort of underlying mechanisms, even when they lack them. We can't help but think about machines in cognitive terms ("It thinks I deleted my file"), no matter how simpleminded the rules are that the machines might actually be following. But inferences that are valid when applied to human beings can be entirely off base when

applied to AI programs. In homage to a central principle of social psychology, we call this the *fundamental overattribution error.*

One of the first cases of this error happened in the mid-1960s, when a chatbot called Eliza convinced some people that it understood what people were saying to it. In fact, Eliza did little more than match keywords, echo the last thing said, and, when lost, reach for a standard conversational gambit ("Tell me about your childhood"). If you mentioned your mother, it would ask you about your family, even though it had no idea what a family really is, or why one would matter. It was a set of tricks, not a demonstration of genuine intelligence.

Despite Eliza's paper-thin understanding of people, many users were fooled. Some users typed away on a keyboard chatting with Eliza for hours, misinterpreting Eliza's simple tricks for helpful, sympathetic feedback. In the words of Eliza's creator, Joseph Weizenbaum:

> People who knew very well that they were conversing with a machine soon forgot the fact, just as theatergoers, in the grip of suspended disbelief, soon forget that the action they are witnessing is not "real." They would often demand to be permitted to converse with the system in private, and would, after conversing with it for a time, insist, in spite of my explanations, that the machine really understood them.

In other cases, overattribution can literally be deadly. In 2016, a Tesla owner came to trust the Autopilot with his life, to the point that he (allegedly) watched *Harry Potter* while the car chauffeured him around. All was well—until it wasn't. After driving safely for hundreds or thousands of miles, the car literally ran into an unexpected circumstance: a white tractor trailer crossed a highway and the Tesla drove directly underneath the trailer, killing the car's owner. (The car appears to have warned him several times to keep his hands on the wheel, but the driver was presumably too disengaged to respond quickly.) The moral of this story is clear: just because something

manages to appear intelligent for a moment or two doesn't mean that it really is, or that it can handle the full range of circumstances a human would.

The second challenge we call the *illusory progress gap:* mistaking progress in AI on easy problems for progress on hard problems. That is what happened with IBM's overpromising about Watson, when progress on *Jeopardy!* was taken to be a bigger step toward understanding language than it really was.

It is possible that DeepMind's AlphaGo could follow a similar path. Go and chess are games of "perfect information"—both players can see the entire board at any moment. In most real-world contexts, nobody knows anything with complete certainty; our data is often noisy and incomplete. Even in the simplest cases, there's plenty of uncertainty; when we decide whether to walk to our doctor's office or take the subway on an overcast day, we don't know exactly how long it will take for the subway to come, or whether it will get stuck, or whether we will be packed in like sardines, or whether we will get soaked if we walk, nor exactly how our doctor will react if we are late. We work with what we've got. Playing Go with yourself a million times, as DeepMind's AlphaGo did, is predictable by comparison. It never had to face uncertainty, or incomplete information, let alone the complexities of human interaction.

There is another way in which games like Go differ deeply from the real world, and it has to do with data: games can be perfectly simulated, so AI systems that play them can collect vast amounts of data cheaply. In Go, a machine can simulate play with humans simply by playing itself; if a system needs billions of data points, it can play itself as often as required. Programmers can get perfectly clean simulation data at essentially no cost. In contrast, in the real world, perfectly clean simulation data doesn't exist, and it's not always possible to collect gigabytes of clean, relevant data by trial and error. In reality, we only get to try out our strategies a handful of times; it's not an option to go to the doctor's office 10 million times, slowly adjusting our parameters with each visit, in order to improve our decisions. If programmers want to train an elder-care robot to help

lift infirm people into bed, every data point will cost real money and real human time; there is no way to gather all the data in perfectly reliable simulations. Even crash-test dummies are no substitute. One has to collect data from actual, squirmy people, in different kinds of beds, in different kinds of pajamas, in different kinds of homes, and one can't afford to make mistakes; dropping people a few inches short of the bed would be a disaster. Actual lives are at stake.* As IBM has discovered not once, but twice, first with chess and later with *Jeopardy!*, success in closed-world tasks just doesn't guarantee success in the open-ended world.

The third contributor to the AI Chasm is what we call the *robustness gap*. Time and again we have seen that once people in AI find a solution that works some of the time, they assume that with a little more work (and a little more data) it will work all of the time. And that's just not necessarily so.

Take driverless cars. It's comparatively easy to create a *demo* of a driverless car that keeps to a lane correctly on a quiet road; people have been able to do that for years. It appears to be vastly harder to make them work under circumstances that are challenging or unexpected. As Missy Cummings, director of Duke University's Humans and Autonomy Laboratory (and former U.S. Navy fighter pilot), put it to us in an email, the issue isn't even how many miles a given driverless car might go without an accident, it's how *adaptable* those cars are. In her words, today's semi-autonomous vehicles "typically perform only under extremely narrow conditions which tell you nothing about how they might perform [under] different operating

* Some initial progress has been made here, using narrow AI techniques. AIs have been developed that play more or less as well as the best humans at the video games *Dota 2* and *Starcraft 2*, both of which show only part of the game world to the player at any given moment, and thus involve a form of the "fog of war" challenge. But the systems are narrow and brittle; for example, AlphaStar, which plays *Starcraft 2*, was trained on one particular "race" of character and almost none of its training would carry over to another race. Certainly there is no reason to think that the techniques used in these programs generalize well in complex real-world situations.

environments and conditions." Being almost perfectly reliable across millions of test miles in Phoenix doesn't mean it is going to function well during a monsoon in Bombay.

This confusion—between how autonomous vehicles do in ideal situations (like sunny days on country roads) and what they might do in extreme ones—could well make the difference between success and failure in the entire industry. With so little attention paid to extreme conditions, and so little in the way of methodology to guarantee performance in conditions that are only starting to be examined, it's quite possible that billions of dollars are being wasted on techniques for building driverless cars that simply aren't robust enough to get us to human-grade reliability. We may need altogether different techniques to achieve the final bit of reliability we require.

And cars are just one example. By and large, in contemporary research in AI, robustness has been underemphasized, in part because most current AI effort goes into problems that have a high tolerance for error, such as ad recommendation and product recommendation. If we recommend five products to you and you like only three of them, no harm done. But in many of the most important

A GROUP OF YOUNG PEOPLE PLAYING A GAME OF FRISBEE

A plausible caption, generated automatically by AI

possible. Hubert Dreyfus once wrote a book about what he thought AI couldn't do—ever. Our book isn't like that. It's partly about what AI can't do now—and why that matters—but it's also about what we might do to improve a field that is still struggling. We don't want AI to disappear; we want to see it improve, radically, such that we can truly count on it to solve our problems. We do have plenty of challenging things to say about the current state of AI, but our criticism is tough love, not a call for anyone to give up.

In short, it is our belief that AI truly can transform the world in important ways, but also that many basic assumptions need to change before real progress can be made. *Rebooting AI* is not an argument to shut the field down (though some may read it that way), but rather a diagnosis of where we are stuck—and a prescription for how we might do better.

The best way forward, we will suggest, may be to look inward, toward the structure of our own minds. Truly intelligent machines need not be exact replicas of human beings, but anyone who looks honestly at AI will see that AI still has a lot to learn from people, particularly from small children, who in many ways far outstrip machines in their capacity to absorb and understand new concepts. Pundits often write about computers being "superhuman" in one respect or another, but in five fundamental ways, our human brains still vastly outperform our silicon counterparts: we can understand language, we can understand the world, we can adapt flexibly to new circumstances, we can learn new things quickly (even without gobs of data), and we can reason in the face of incomplete and even inconsistent information. On all of these fronts, current AI systems are non-starters. We will also suggest that a current obsession with building "blank slate" machines that learn everything from scratch, driven purely from data rather than knowledge, is a serious error.

If we want machines to reason, understand language, and comprehend the world, learning efficiently, and with human-like flexibility, we may well need to first understand how humans manage to do

so, and to better understand what it is that our minds are even trying to do (hint: it's not all about the kind of correlation-seeking that deep learning excels at). Perhaps it is only then, by meeting these challenges head-on, that we can get the reboot that AI so desperately needs, and create AI systems that are deep, reliable, and trustworthy.

In a world in which AI will soon be as ubiquitous as electricity, nothing could be more important.

2

What's at Stake

A lot can go wrong when we put blind faith in big data.
—CATHY O'NEIL, TED TALK, 2017

On March 23, 2016, Microsoft released Tay, designed to be an exciting and new chatbot, not hand-wired entirely in advance, like the original chatbot, Eliza, but instead developed largely by learning from user interactions. An earlier project, Xiaoice, which chats in Chinese, had been a huge success in China, and Microsoft had high hopes.

Less than a day later the project was canceled. A nasty group of users tried to drown Tay in racist, sexist, and anti-Semitic hate. Vile speech in, vile speech out; poor Tay was posting tweets like "I fucking hate the feminists" and "Hitler was right: I hate the Jews."

Elsewhere on the internet, there are all sorts of problems great and small. One can read about Alexas that spooked their owners with random giggles, iPhone face-recognition systems that confused mother and son, and the Poopocalypse, the Jackson Pollack–like mess that has happened more than once when a Roomba collided with dog waste.

More seriously, there are hate speech detectors that get easily fooled, job candidate systems that perpetuate bias, and web browsers and recommendation engines powered by AI tools that have been tricked into pushing people toward ludicrous conspiracy theories. In China, a face-recognition system used by police sent a jaywalking ticket to an innocent person who happened to be a well-known entrepreneur when it saw her picture on the side of a bus, not realizing that a larger-than-life photo on a moving bus was not the same thing as the entrepreneur herself. A Tesla, apparently in "Summon"

mode, crashed while backing out of its owners' garage. And more than once, robotic lawnmowers have maimed or slaughtered hedgehogs. The AI that we have now simply can't be trusted. Although it often does the right thing, we can never know when it is going to surprise us with errors that are nonsensical or even dangerous.

And the more authority we give them, the more worried we should be. Some glitches are mild, like an Alexa that randomly giggles (or wakes you in the middle of the night, as happened to one of us) or an iPhone that autocorrects what was meant as "Happy Birthday, dear Theodore" into "Happy Birthday, dead Theodore." But others—like algorithms that promote fake news or bias against job applicants—can be serious problems. A report from the group AI Now has detailed many such issues in AI systems in a wide variety of applications, including Medicaid eligibility determination, jail term sentencing, and teacher evaluation. Flash crashes on Wall Street have caused temporary stock market drops, and there have been frightening privacy invasions (like the time an Alexa recorded a conversation and inadvertently sent it to a random person on the owner's contact list); and multiple automobile crashes, some fatal. We wouldn't be surprised to see a major AI-driven malfunction in an electrical grid. If this occurs in the heat of summer or the dead of winter, a large number of people could die.

Which is not to say that we should stay up nights worrying about a Skynet-like world in which robots face off against people, at least any time in the foreseeable future. Robots don't yet have the intelligence or manual dexterity to reliably navigate the world, except in carefully controlled environments. Because their cognitive faculties are so narrow and limited, there is no end to the ways they can be stymied.

More important, there is no reason to think that the robots will, science fiction style, rise up against us. After sixty years of AI, there is not the slightest hint of malice; machines have demonstrated zero interest in tangling with humans for territory, possessions, bragging