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. Copyright © 2002 by Wislawa Szymborska. English translation copyright © 2006 by Houghton Mifflin Harcourt Publishing Company. Reprinted by permission of Houghton Mifflin Harcourt Publishing Company. All rights reserved.

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

Ebook ISBN 9781524748265

www.pantheonbooks.com

5

Where's Rosie?

> In ten years' time Rossum's Universal Robots will be making so much wheat, so much material, so much of everything that nothing will cost anything.
>
> —KAREL ČAPEK, WHO COINED THE WORD "ROBOT," IN *R.U.R.,* THE 1920 PLAY THAT INTRODUCED THEM

> We are still in the infancy of having real autonomous interacting, learning, responsible, useful robots in our environment.
>
> —MANUELA VELOSO, "THE INCREASINGLY FASCINATING OPPORTUNITY FOR HUMAN-ROBOT-AI INTERACTION: THE COBOT MOBILE SERVICE ROBOTS," APRIL 2018

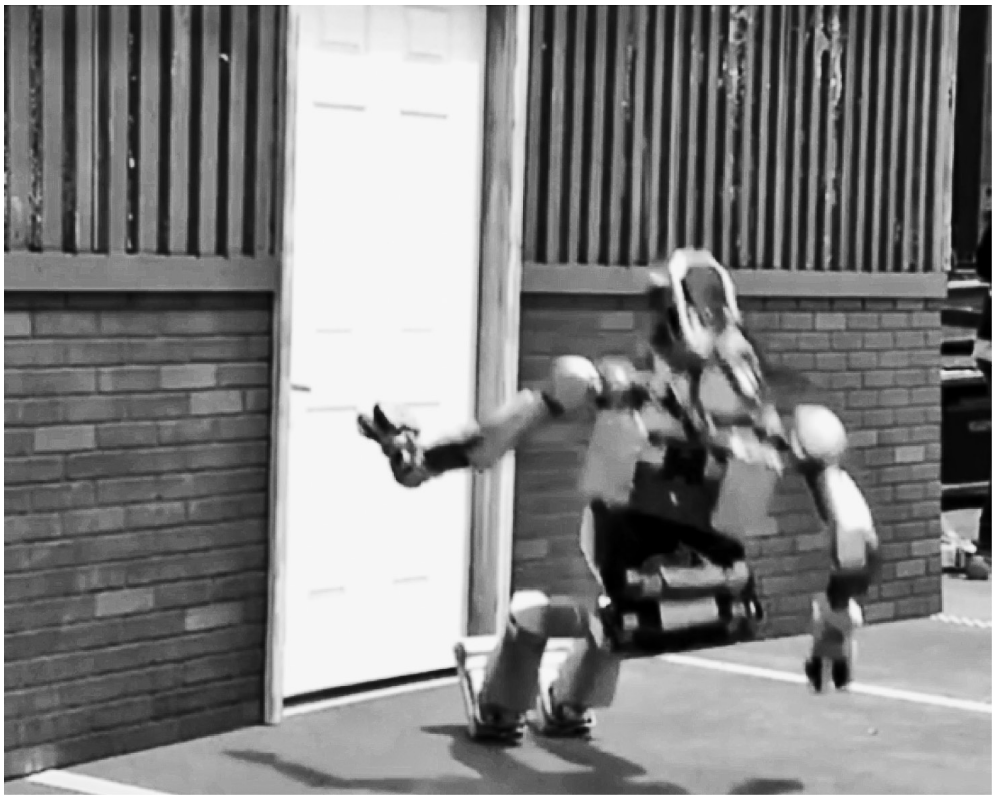
Worried about superintelligent robots rising up and attacking us?

Don't be. At least for now, here are six things you can do in the event of a robot attack.

- Close your doors, and for good measure, lock them. Contemporary robots struggle greatly with doorknobs, sometimes even falling over as they try to open them. (In fairness, we've seen one demo that shows a robot opening one particular doorknob, in one particular lighting condition, but the AI probably doesn't generalize. We don't know of any demos that show that robots can open a broad range of doorknobs of different shapes and sizes, let alone in different

lighting conditions, or any demo that shows a robot opening a locked door—even with a key.)

- Still worried? Paint your doorknob black, against a black background, which will greatly reduce the chance that the robot will even be able to see it.



(Source: IEEE Spectrum)

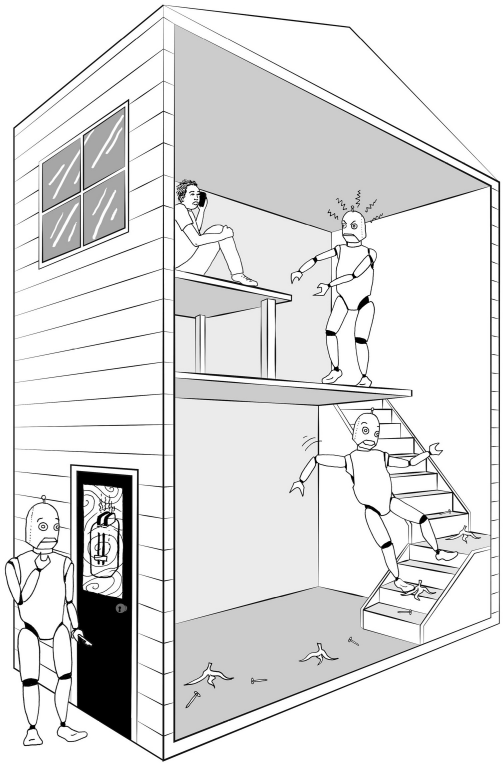
Robot falling backward trying to open a door

- For good measure, put a big poster of a school bus or a toaster on your front door (see chapter 3). Or wear a T-shirt with a picture of a cute baby. The robot will be totally confused, think you are a baby, and leave you alone.
- If that doesn't work, go upstairs, and leave a trail of banana peels and nails along the way; few robots will be able to handle an impromptu obstacle course.
- Even the ones that climb stairs probably can't hop up onto a table, unless they have been specifically trained for the task.

You probably can, so hop up on the table, or climb a tree, and call 911.

- Relax; either 911 will arrive or the robot's battery will soon run out. Free-ranging robots currently typically last for a few hours, but not much more between charges, because the computers inside them demand such vast amounts of energy.

OK, this might be facetious, and maybe someday robots will be able to crash through doors and hop on tables, but for now the robots we know about are easily confused. At least in the near term, we needn't worry about Skynet, or even that robots will take our jobs.



Foiling a robot attack

To the contrary our biggest fear is that the robot revolution will be stillborn, because of an unreasonable fear of unlikely things.

In the movies, robots are often cast either as heroes or as demons; R2-D2 frequently rushes in to save the day; while the Terminator is here to slaughter us all. Robots want either to please their owners or to annihilate them. In the real world, robots generally don't have personalities or desires. They aren't here to slaughter us or to take our land, and they certainly don't have the wherewithal to save us from a dark lord. They don't even blip all that much, like R2-D2. Instead, for the most part, they hide on assembly lines, doing dull tasks humans would never want to do.

And robot companies, for the most part, aren't much more ambitious. One company we talked to recently is focused on building robots for excavating the foundations of buildings, another is focused on picking apples. Both seem like good business propositions, but they are not exactly the sort of stuff we dreamed of when we were kids. What we really want is Rosie, the all-purpose domestic robot from the 1960s television show *The Jetsons,* who could take care of everything in our house—the plants, the cats, the dishes, and the kids. Oh, to never have to clean anything again. But we can't buy Rosie, or anything like it, for love or for money. As we write this, there are rumors that Amazon might roll out a version of Alexa that roams around on wheels, but that's still a long way from Rosie.

The truth is that for now the bestselling robot of all time isn't a driverless car or some sort of primitive version of C-3PO, it's Roomba, that vacuum-cleaning hockey puck of modest ambition, with no hands, no feet, and remarkably little brain that we mentioned in the opening chapter, about as far from Rosie the Robot as we can possibly imagine.

To be sure, pet-like home robots are already available, and "driverless" suitcases that follow their owners around through airports may come soon enough. But the chance that a robot will be cooking and cleaning and changing kids' diapers before 2025 is practically nil. Outside of factories and warehouses, robots are still a curiosity.[*1]

What would it take to get from the modestly impressive but still greatly limited Roomba to a full-service humanoid companion like C-3PO or Rosie that could simplify almost every aspect of our domestic lives and transform the lives of the elderly and disabled, and literally save all of us hours of labor every week?

To begin with, it's important to realize that Roomba is a very different sort of creature. Inventor Rodney Brooks's great insight—inspired by reflecting on how insects with tiny brains could do complex things like flying—was that Roomba doesn't need to be very smart. Vacuuming is a mundane job and can be done decently well (not perfectly) with only a modest bit of intelligence. Even with tiny amounts of computer hardware you could make a robot that could do something useful, that people would want to spend real money on—so long as you kept the task narrow enough. If you want to pick up most of the dust, most of the time, in an ordinary room, you can just go back and forth, spiraling around and changing direction every now and then when you bump into something. It's often pretty inefficient, going over the same parts of the floor many times. But most of the time, if it doesn't miss some part of a room that can only be reached through a narrow passageway, it gets the job done.

The real challenge is to go beyond vacuuming and build robots that can carry out a broad range of complex physical tasks that we humans accomplish in daily life, from opening vacuum-sealed jars, twist-off bottle caps, and envelopes, to weeding, hedge-clipping, and mowing the lawn, to wrapping presents, painting walls, and setting the table.

Of course, there has been some progress. Our good friend the roboticist Manuela Veloso has built robots that can safely wander the halls of Carnegie Mellon University, and we have seen demos of robots lifting vastly more than their own body weight. Autonomous drones (which are flying robots) can already do some amazing things, like tracking runners as they jog along mountain trails, and

(in the case of Skydio's self-flying camera) automatically avoiding trees along the way.

If you spend a few hours watching YouTube, you can see dozens of demos of robots that (at least in videos) seem vastly more powerful than Roomba. But the key word is "demo." None of it is ready for prime time. In 2016, Elon Musk announced plans to build a robotic butler, but so far as we can tell, there hasn't been much progress toward that goal. Nothing currently available commercially feels like a breakthrough, with the possible exception of the aforementioned recreational drones, which are thrilling (and tremendously useful for film crews), but not exactly Rosie either. Drones don't need to pick up things, manipulate them, or climb stairs; aside from flying around and taking photos, they aren't called on to do very much. SpotMini, a sort of headless robotic dog, is supposed to be released soon, but it remains to be seen what it will cost, and what it will be used for. Boston Dynamics' Atlas robot, a humanoid robot, about five feet tall, and 150 pounds, can do backflips and parkour, but that parkour video you saw on the web? It took twenty-one takes, in a carefully designed room; you shouldn't expect that it would be able to do the same at the playground with your kids.

Even so, there's lot of exciting hardware on the way. In addition to SpotMini and Atlas, both of which are amazing, Boston Dynamics' robots include WildCat, "the world's fastest quadruped robot," that can gallop at twenty miles per hour; and BigDog, "the First Advanced Rough-Terrain Robot," which stands three feet high and weighs 240 pounds, can run at seven miles an hour, climb slopes up to 35 degrees, walk across rubble, traverse muddy hiking trails, wade through snow and water, and carry a 100-pound payload. And of course every putative driverless car is just a robot in a car package. (And for that matter, submersibles, like Alvin, are robots, too, and so are the Mars Rovers.) Other researchers, like MIT's Sangbae Kim, are working on impressively agile hardware as well. All of this costs way too much now for the home, but someday prices will come down and robots might be in most homes.

Perhaps the most important use of robots to date has been in the shutdown and cleanup of the Fukushima nuclear reactor, after it was destroyed in the 2011 tsunami. Robots from the iRobot company were sent into the reactor to determine the state of things inside, and they continue to be used in cleaning and maintenance of the site. Though the robots were mostly controlled via radio communication by human operators outside, they also had important, though limited, AI capabilities: they could build maps, plan optimal paths, right themselves should they tumble down a slope, and retrace their path when they lost contact with their human operators.

The real issue is software. Driverless cars can propel themselves, but not safely. SpotMini is capable of amazing feats, but thus far it has been mostly teleoperated, which is to say that someone with a joystick is offstage telling the robot what to do. To be sure, the mechanical and electrical engineers and the materials scientists who make robots will be kept busy for years—there is still a long way to go in making better batteries, improving affordability, and building bodies that are sufficiently strong and dexterous—but the real bottleneck is getting robots to do what they do safely and autonomously.

What will it take to get there?

In *Star Trek: The Next Generation,* the answer is simple: all you need is what Lieutenant Commander Data has: a "positronic brain." Unfortunately, we are still not exactly sure what that is, how it might work, or where we might order one.

In the meantime, there are several things that one can expect of virtually any intelligent creature—robot, human, or animal—that aspires to be more sophisticated than Roomba. To begin with, any intelligent creature needs to compute five basic things: where it is, what is happening in the world around it, what it should do right now, how it should implement its plan, and what it should plan to do over the longer term in order to achieve the goals it has been given.

In a less-sophisticated robot focused on a single task, it might be possible to sidestep these computations to some degree. The original Roomba model had no idea of where it was, it didn't keep track of the map of the territory it had navigated, and it made no plans; it knew little more than whether or not it was moving and whether it had bumped into something recently. (More recent Roomba models do build maps, in part so that they can be more efficient, in part to make sure that they don't miss spots via an otherwise largely random search.) The question of what to do now never arose; its only goal was to vacuum.

But Roomba's elegant simplicity can only go so far. In the daily life of a more full-service domestic robot, many more choices would arise, and decision-making would thus become a more complex process—and one that would depend on having a vastly more sophisticated understanding of the world. Goals and plans could easily change from one moment to the next. A robot's owner might instruct it to unload the dishwasher, but a good domestic robot wouldn't forge ahead, no matter what; it would adapt when circumstances change.

If a glass plate falls and breaks on the floor next to the dishwasher, a robot might need to find another route to the dishwasher (a change to its short-term plans), or, better yet, it might realize that its priority is to clean up the broken glass and put the dishes on hold while it does that.

If the food on the stove catches fire, the robot has to postpone unloading the dishwasher until it has put the fire out. Poor Roomba would keep vacuuming in the middle of a Category 5 hurricane. We expect more of Rosie.

Precisely because the world changes constantly, fixed answers to the core questions about goals, plans, and the environment will never do. Instead, a high-quality domestic robot will constantly need to reevaluate. "Where am I?," "What is my current status?," "What risks and opportunities are there in my current situation?," "What should I be doing, in the near term and the long term?," and "How should I execute my plans?"[*2] Each of these questions must be continuously

addressed in a constant cycle, a robotic counterpart to the so-called OODA loop introduced by the legendary air force pilot and military strategist John Boyd: observe, orient, decide, and act.

The good news is that over the years, the field of robotics has gotten pretty good at implementing some parts of the robot's cognitive cycle. The bad news is that most others have seen almost no progress.

Let's start with the success stories: localization and motor control.

Localization is harder than you might think. The obvious way to start is with GPS. But GPS has until recently only been accurate to within about ten feet, and it doesn't work particularly well indoors. If that were all our hypothetical domestic robot had to work with, it could easily think it was in the bathroom when it was really on the staircase.

Military and specialized GPS can be much more accurate, but isn't likely to be available to consumer robots, which means consumer robots can't rely just on GPS. Luckily, robots can use many clues to figure out where they are, such as dead reckoning (which tracks a robot's wheels to estimate how far it has gone), vision (a bathroom looks very different from a staircase), and maps (which might be constructed in a variety of ways). Over the years, roboticists have developed a family of techniques called SLAM, short for Simultaneous Localization And Mapping, which allows robots to put together a map of their environment and to keep track of where they are in the map and where they are headed. At each step, the robot goes through the following steps:

- The robot uses its sensors to see the part of its environment that is visible from its current position.
- It improves its current estimate of its position and orientation by matching what it is seeing against objects in its mental map.

- It adds to its mental map any objects, or parts of objects, that it has not seen before.
- It either moves (generally forward) or turns, and adjusts its estimate of its new position and orientation by taking into account how much it has moved or turned.

Although no technique is perfect, SLAM works well enough that you can plop a robot down at a random place in a well-mapped building and expect it to figure out where it is and, in conjunction with other software, how to get where it needs to go. It also allows robots to construct maps as they explore a space. Orientation, in Boyd's sense, is more or less a solved problem.

Another area with considerable progress is often called "motor control": the job of guiding a robot's motions, such as walking, lifting things, rotating its hands, turning its head, or climbing stairs.

For driverless cars, the motor control side of what needs to be done is comparatively simple. A car has only limited options, revolving around the gas pedal, the brakes, and the steering wheel. An autonomous vehicle can change its speed (or stop) and it can change its heading, by steering. There's not much else to be calculated, on the control side. Unless the car can fly, it doesn't even need to worry about the z-coordinate of going up or down in space. Computing the desired states for the steering wheel, brakes, and gas pedal from a desired trajectory is straightforward math.

The situation is way more complicated in a humanoid (or animal-like, or insect-like) robot, with multiple joints that can be moved in many ways. Suppose that there is a cup of tea on a table, and a humanoid robot is supposed to reach out its arm and grasp the cup's handle between two fingers. First, the robot has to figure out how to move the various parts of its arm and hand so that they end up at the right place without, in between, bumping into the table, having one part bump into another, or knocking the teacup over. Then it has to

exert enough force on the teacup handle that it has a firm grasp but not so much force that it shatters the china. The robot must calculate a path between where it wants to go given where it is and the obstacles in its way, and then devise a complex plan (effectively a mini computer program, or a custom-built neural network) that specifies the angle at the joints and the force at the joints between body parts, and how they should change over time, perhaps as a function of feedback, in a way that never allows the contents to spill. Even just in reaching for the teacup, there might be five or more joints involved—the shoulder, the elbow, the wrist, and two fingers, with many complex interactions between them.

Despite the complexity of the problem, in recent years there has been considerable progress, most visibly at Boston Dynamics, the robot company that we mentioned earlier, which is run by Marc Raibert, a researcher with deep training in the problem of human and animal motor control. Drawing on this expertise, Raibert's robots, such as BigDog and SpotMini, move much like animals. Their software rapidly and continuously updates the forces in the actuators (the robot's "muscles") and integrates that with feedback from the robot's sensors, so that they can dynamically replan what they are supposed to do, as they are doing it (rather than just planning everything in advance and hoping for the best). Raibert's team has been able to do a bunch of things that used to be pretty challenging for many robots, like walking on uneven surfaces, climbing stairs, and even resisting forces that might otherwise knock a less stable robot down.

Lots of labs at places like Berkeley and MIT are making good progress on motor control too. YouTube has videos of laboratory demonstrations of robots that open doors, climb stairs, toss pizzas, and fold towels, although typically in carefully controlled circumstances. Although the motor control of humans remains more versatile, particularly when it comes to manipulating small objects, robots are catching up.

Then again, most of what we can glean about the current state of motor control in robotics we get from demo videos, and videos are

often misleading. Often the videos have been sped up, implicitly equating what a robot could do in a minute or an hour with what a person could do in a few seconds, and sometimes relying on human operators behind the scenes. Such demo videos are proofs of concept, frequently representing the best case of something that is not truly stable, not products that are ready for shipping. They prove that a robot can, in principle, given enough time, eventually be programmed to do the physical aspects of many tasks. But they don't always tell us whether various tasks can be performed efficiently or—most important—autonomously, which of course is the ultimate goal. Eventually you should be able to say to a robot "Clean my house," and after a little bit of training, it should not only vacuum, but also dust, clean windows, sweep the porch, straighten out the books, toss the junk mail, fold the laundry, take out the garbage, and load the dishwasher. What the demos show is that we now have the hardware to do at least some of these tasks; the physical aspects of the jobs won't be the rate-limiting steps. The real challenge will be on the mental side, of having the robot correctly interpret your possibly ambiguous and vague request—relative to human goals—and coordinate all of its plans in a dynamically changing world.

As with AI more generally, the biggest challenge will be robustness. In almost any demo you watch, you see a robot doing something in the most ideal circumstances imaginable, not in a cluttered and complex environment. If you look carefully at videos of robots folding towels, you will find that the laundry is always some bright color, and the background some dark color, in an otherwise empty room, making it easier for the computer software to separate the towels from the rest of the room. In an actual home with dim light and towels that blend in with the background, pandemonium might ensue, when the robot occasionally mistakes bits of wall for bits of towel. A robotic pancake flipper might work fine in a restaurant in which it could be housed in a room with blank walls, but have trouble in a cluttered bachelor pad, where stacks of unread mail might inadvertently wind up flipped, fried, and in flames.

Motor control in the real world is not just about controlling the actions of limbs and wheels and so forth in the abstract. It's about controlling those actions relative to what an organism perceives, and coping when the world is not exactly as anticipated.

Situational awareness is about knowing what could happen next. Is a storm coming? Could that pot on the stove catch fire if I forget to turn it off? Could that chair be about to fall over? (Parents of toddlers tend to have heightened awareness of the latter.) One aspect of situational awareness is about looking for risk, but it can be about looking for opportunity or reward, too. For instance, a driverless car might notice that a new shortcut has opened up, or an unexpected parking spot is vacant. A home robot that was trying to unclog a drain could discover a new use for a turkey baster. On a well-controlled factory floor, situational awareness can similarly be a relatively manageable problem, limited to questions like "Is there an obstacle here?" and "Is the conveyor belt running?"

In the home, on the other hand, situations, and the risks, rewards, and opportunities that go with them, can be vastly more challenging and more complex. Sitting in your living room, you might have literally hundreds of options, and thousands of parameters could change the nature of the situation. You can get up, walk to the dining room or the kitchen, or turn on the television, or a pick up a book, or tidy the coffee table. Any of those might seem like reasonable activities on an ordinary day, but not if the smoke detector goes off or a hurricane approaches.

In calculating what's going on, and the risks and opportunities in any given moment, you (as a person) constantly combine sight with smell and hearing (and perhaps touch and taste) and a sense of where your own body is, along with an awareness of the other beings that might be in the room, your overall goals (what are you trying to do that hour? that day? that month?), and hundreds of other variables (Is it raining? Have I left a window open? Could an insect

or animal wander in uninvited?). If assembly lines are closed worlds, homes are about as open-ended as you can get, and hence a serious challenge for robotics.

Driverless cars are somewhere in between. The vast majority of the time, figuring out what is going on mainly requires computing a few things: Which way am I going and how fast? Which way is the road turning? What other objects are nearby? Where are they and how are they moving? (which you can compute by comparing data from different time steps); and Where may I drive (e.g., where are the lanes, or opportunities to make a turn)? But all bets may be off in a tornado, an earthquake, or a fire, or even if there is a fender-bender or a toddler in a Halloween costume that diverts traffic.

The part of situational awareness that is fairly well handled by current AI is the job of identifying objects in some surroundings: simple object recognition is deep learning's forte. Machine-learning algorithms can now, with some degree of accuracy, identify basic elements in many scenes, from tables and pillows in a home to cars on the road. Even in simple identification, though, there are some serious problems; few object-recognition systems are robust enough to notice changes in lighting, and the more cluttered a room is, the more likely they are to grow confused. And it's not enough to note that there is a gun somewhere in the image; it is important to know whether the gun is on the wall as part of a painting (in which case it can be safely ignored), or a real object on a table, or in someone's hands pointed at someone. Even more than that, simple object-recognition systems fall far short of understanding the *relations* between objects in a scene: a mouse *in* a trap is very different from a mouse *near* a trap; a man riding a horse is very different from a man carrying a horse.

But labeling the objects in a scene is less than half the battle. The real challenge of situational awareness is to understand what all those objects collectively mean; to our knowledge there has been little or no research on this problem, which is clearly vastly harder. We don't know of any current algorithm, for example, that could look at two different scenes in which there was a fire in a living room, and

reliably realize that in one case the fire is in a fireplace giving delightful warmth on a wintry day, and in the other case you had better put the fire out as fast as you can and/or call the fire department. To even approach the problem within the dominant paradigm one would probably need a bunch of labeled data sets for different kinds of homes (wooden, concrete, etc.) and different kinds of fires (grease, electrical, etc.); nobody has a general purpose system for understanding fire.

And the changing nature of the world makes situational awareness even harder; you don't want to look at the world as a snapshot, you want to see it as an unfolding movie, to distinguish objects that are toppling over from objects that are stable, and to distinguish cars pulling into a parking spot from cars pulling out of a parking spot.

Making all this even more challenging is the fact that the robot itself is both changing (for example, as it maneuvers around) and is an agent of other changes, which means that a robot must predict not only the nature of the world around it but the consequences of its own actions. On a factory floor, where everything is tightly controlled, this may be fairly easy; either the car door becomes securely attached to the car chassis, or it doesn't. In an open-ended environment, prediction becomes a real challenge: If I'm looking for the coffee, should I open the cabinet? Should I open the fridge? Should I open the mayonnaise jar? If I can't find the cover for the blender, is it OK to run the blender without it? Or to cover the blender with a plate? Even the factory floor becomes challenging the minute there is a loose screw in an unexpected place. Elon Musk blamed initial struggles in manufacturing the Tesla Model 3 on "too much automation." We suspect that a large part of the problem was that the process and environment for building the cars was dynamically changing, and the robots couldn't keep up because their programming wasn't flexible enough.

Some of this can be discovered by world experience, but the consequences of putting the cat in the blender shouldn't be something an AI learns through trial and error. The more you can

make solid inferences without trying things out, the better. In this kind of everyday reasoning, humans are miles and miles ahead of anything we have ever seen in AI.

Perhaps an even bigger unsolved challenge lies in figuring out what is the best thing to be doing at any given moment, which is much harder (from a programming perspective) than one might initially realize.

To better understand what challenges our hypothetical domestic robot might face, let's consider three concrete scenarios, typical of what we might ask of it.

First scenario: Elon Musk is giving an evening party, and he wants a robot butler to go around serving drinks and hors d'oeuvres. For the most part this is straightforward: the robot moves around carrying plates of drinks and snacks; it collects empty glasses and plates from the guests; if a guest asks for a drink, the robot can bring it. At first blush, this might not seem far away. After all, years ago the now-defunct robot company Willow Garage had a demo of their humanoid prototype robot PR2 fetching a beer from a refrigerator.

But just as with driverless cars, true success lies in getting the details right. Real homes and real guests are complicated and unpredictable. The PR2 beer run was carefully constructed. There were no dogs, no cats, no broken bottles, and no children's toys left on the floor. Even the fridge was specially arranged, according to a colleague of ours, in ways that made the beer itself particularly accessible. But in the real world, any number of unexpected things, big and small, can go wrong. If the robot goes into a kitchen to get a wine glass and finds a cockroach in the glass, it has to form a plan that it may never have executed before, perhaps dumping the roach out of the glass, rinsing and refilling it. Or maybe the butler robot may discover a crack in the glass, in which case the glass should be safely disposed of. But what is the likelihood that some programmer is going to anticipate that exact contingency, when all of Apple's best

iPhone programmers still can't even reliably automate the process of creating a calendar entry based on text in your email?

The list of contingencies is essentially endless—the Achilles' heel of narrow AI. If the robotic butler sees that a cracker has fallen on the floor, it needs to figure out how to pick up the cracker and throw it out without disturbing the guests, or it needs to be able to predict that picking up the cracker in a crowded room isn't worth the trouble, because it would cause too much commotion. No simple policy will do here, though. If the robot sees an expensive earring on the floor, rather than a cracker, the balance of the equation changes; it may be worth rescuing the earring regardless of the commotion.

Most of the time, the robot should do humans no harm. But what if a drunk guy is walking backward, oblivious to an infant crawling behind him? The robot butler ought to interfere at this point, perhaps even grabbing the drunk adult, to protect the child.

So many things might happen that they cannot possibly all be enumerated in advance, and they cannot all be found in any data set used for training. A robot butler is going to have to reason, predict, and anticipate, on its own, and it can hardly go crying to human "crowd workers" all night long every time a small decision is to be made. Surviving a night at Elon's mansion would be, from a cognitive perspective, a pretty major task.

Of course we can't all afford robotic butlers, at least not until the price comes down by a factor of a million. But now let's consider a second scenario, far less frivolous: robotic companions for the elderly and the disabled. Suppose Blake is recently blind and he would like his companion robot to help him go grocery shopping. Again, so much easier said than done, because so many kinds of things can happen. To begin with, there is basic navigation. On the way to the grocery store, Blake's companion robot will need to navigate unexpected obstacles of all kinds.

Along the way, they might encounter curbs, puddles, potholes, police, pedestrians lost in their phones, and children weaving about on scooters and skateboards. Once in the store they may need to navigate narrow aisles, or temporary taster stands that subtly alter

the functional layout of the grocery store, to say nothing of the inventory people or the people cleaning the floor after somebody accidentally dropped a jar of jam. The companion robot will have to guide Blake around or through these, in addition to finding its own way. Meanwhile Blake may be accosted by an old friend, a helpful stranger, a panhandler, a policeman, a friendly dog, an unfriendly dog, or a mugger; each has to be recognized and dealt with in a different way. At the store, things have to be reached and grasped (in different ways for different items, red peppers differently from cereal boxes differently from pints of ice cream), and put into the shopping basket without cracking the eggs or piling the soup cans on top of the bananas. The shopping basket itself needs to be recognized, even though these vary in shape and size from store to store; likewise the means of payment and details of how groceries are bagged vary from one store to the next. Thousands of contingencies, varying from one shopping experience to the next, impossible to fully anticipate and program in advance.

As a third scenario, consider something like the Fukushima nuclear disaster. Imagine a building that has partially collapsed in an earthquake, and a nuclear reactor is about to melt down. A rescue robot sent inside a crisis zone has to judge what it can and cannot safely do: Can it break through a door, or cut through a wall, or does that risk further collapse? Can it safely climb a ladder that was designed for humans? If the rescue robot finds someone, what should be done? The person may be able to walk out under their own power once a path is cleared, or they may be pinned and need to be freed; or may be injured and have to be carried out carefully. If there are multiple people, the robot may have to triage, deciding which injuries should be treated first, and which not at all, given limited medical resources. If there is valuable property, the rescue robot should consider the value of the item (is it irreplaceable art?) and the urgency of removing it. All this may require deep understanding of a situation that is incompletely known, unforeseen, with features that may well be unusual or unique.

What's more, a robot needs to consider the dangers of inaction as well as action. A high-quality robot butler ought to be able to spot a Christmas tree that is listing at a dangerous angle, and readjust it, in order to keep the tree from falling over and potentially sparking and then fueling an electrical fire.

None of this is a strong point of current robots, or the AI that drives them.

So here's where things stand today, as we approach the sixty-fifth anniversary of AI: roboticists have done an excellent job of getting robots to figure out where they are, and a fairly good job of figuring how to get robots to perform individual behaviors.

But the field has made much less progress in three other areas that are essential to coping in the open-ended world: assessing situations, predicting the probable future, and deciding, dynamically, as situations change, which of the many possible actions makes the most sense in a given environment.

There is no general purpose solution either to determining what is possible and important in any given scenario, or to figuring out what a robot should do in complex and unpredictable environments. At the present time it is challenging but (with hard work) feasible to get a robot to climb a staircase or walk on uneven ground, as the Boston Dynamics prototypes have shown; it is far harder to leave a robot to clean a kitchen entirely by itself.

In a limited world, one can memorize a large number of contingencies, and interpolate between them to make guesses for unfamiliar scenarios. In a truly open-ended world, there will never be enough data. If the applesauce is growing mold, the robot needs to figure out how to respond, even if the robot has never seen anything like this before. There are just too many possibilities for it to memorize a simple table listing what to do in every circumstance that could arise.[*3]

The real reason we don't have general-purpose domestic robots yet is that we don't know how to build them to be flexible enough to cope with the real world. Because the space of possibilities is both vast and open-ended, solutions that are driven purely by big data and deep learning aren't likely to suffice. Classical AI approaches, too, have been brittle, in their own ways.

All of this points, once again, to the importance of rich cognitive models and deep understanding. Even in the situation of a driverless car, a machine's internal models will need to be richer than what AI typically incorporates. Current systems are mostly limited to identifying common objects such as bicycles, pedestrians, and other moving vehicles. When other kinds of entities enter in, such limited systems can't really cope. For example, as of 2019, Tesla's Autopilot seems to have limited representations of stationary objects, such as stopped fire trucks or billboards (its first fatal accident may have been partly due to its misinterpreting a left-turning truck, with much of its mass higher than a car, as a billboard).

In the case of our household robot, the richness of the underlying cognitive model has to be considerably greater. While there are only a handful of common elements on a highway, in an average living room one might encounter chairs, a sofa or two, a coffee table, a carpet, a TV, lamps, bookcases with books, the fish tank, and the cat, plus a random assortment of children's toys. In a kitchen one might encounter utensils, appliances, cabinets, food, a faucet, a sink, more chairs and tables, the cat bowl, and again the cat. And, of course, even though kitchen utensils are usually found in the kitchen, a knife that finds its way into the living room can still hurt someone.

In many ways, what we see here echoes what we saw in the last chapter, on learning to read. Building a robot is a very different challenge from building a machine that can read, much more physical and much less about narrative and interpretation (and also much more potentially dangerous—spilling boiling tea on someone is

much worse than mucking up a translation of a new story), yet we have converged on the same place.

Just as there can be no reading without rich cognitive models, there can be no safe, reliable domestic robots without rich cognitive models. Along with them, a robot will need a healthy dose of what is colloquially known as common sense: a rich understanding of the world, and how it works, and what can and cannot plausibly happen in various circumstances.

No existing AI system has all that. What sort of intelligent system does have rich cognitive models, and common sense? The human mind.

---

* 1 Fast-food-cooking restaurant bots are another story; in high-volume chains like McDonald's, with great control of their internal environments and high labor costs, increased automation is likely to come soon.

* 2 We are, of course, anthropomorphizing here, in suggesting that the robot has any sense of "I" or asks itself this kind of question. It would be more accurate to write that the robot's algorithm calculates where it is, what is its current status, what are the risks and opportunities, what it should do next, and how it should execute its plans.

* 3 Programs that play games like chess and Go must, of course, deal with situations that they have not seen before, but there the kinds of situations that can arise and the choice of actions can be systematically characterized, and the effects of actions can be reliably predicted, in ways that do not apply in an open-ended world.