LIFE 3.0

-

Being Human in the Age of Artificial Intelligence

Max Tegmark

Alfred A. Knopf
New York
2017

Chapter 3

The Near Future: Breakthroughs, Bugs, Laws, Weapons and Jobs

If we don't change direction soon, we'll end up where we're going.

Irwin Corey

What does it mean to be human in the present day and age? For example, what is it that we really value about ourselves, that makes us different from other life forms and machines? What do other people value about us that makes some of them willing to offer us jobs? Whatever our answers are to these questions at any one time, it's clear that the rise of technology must gradually change them.

Take me, for instance. As a scientist, I take pride in setting my own goals, in using creativity and intuition to tackle a broad range of unsolved problems, and in using language to share what I discover. Fortunately for me, society is willing to pay me to do this as a job. Centuries ago, I might instead, like many others, have built my identity around being a farmer or craftsman, but the growth of technology has since reduced such professions to a tiny fraction of the workforce. This means that it's no longer possible for everyone to build their identity around farming or crafts.

Personally, it doesn't bother me that today's machines outclass me at manual skills such as digging and knitting, since these are neither hobbies of mine nor my sources of income or self-worth. Indeed, any delusions I may have held about my abilities in that regard were crushed at age eight, when my school forced me to take a knitting class which I nearly flunked, and I completed my

project only thanks to a compassionate helper from fifth grade taking pity on me.

But as technology keeps improving, will the rise of AI eventually eclipse also those abilities that provide my current sense of self-worth and value on the job market? Stuart Russell told me that he and many of his fellow AI researchers had recently experienced a “holy shit!” moment, when they witnessed AI doing something they weren’t expecting to see for many years. In that spirit, please let me tell you about a few of my own HS moments, and how I see them as harbingers of human abilities soon to be overtaken.

Breakthroughs

Deep Reinforcement Learning Agents

I experienced one of my major jaw drops in 2014 while watching a video of a DeepMind AI system learning to play computer games. Specifically, the AI was playing Breakout (see figure 3.1), a classic Atari game I remember fondly from my teens. The goal is to maneuver a paddle so as to repeatedly bounce a ball off a brick wall; every time you hit a brick, it disappears and your score increases.

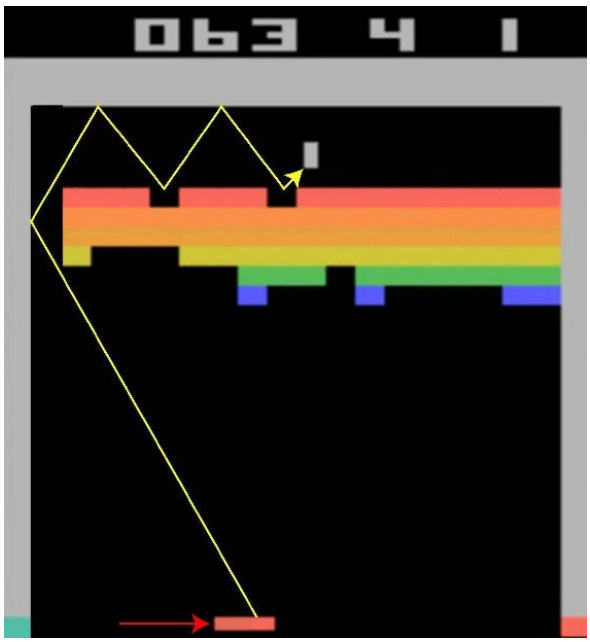

Figure 3.1: After learning to play the Atari game Breakout from scratch, using deep reinforcement learning to maximize the score, the DeepMind AI discovered the optimal strategy: drilling a hole through the leftmost part of the brick wall and letting the ball keep bouncing around behind it, amassing points very rapidly. I've drawn arrows showing the past trajectories of ball and paddle.

I'd written some computer games of my own back in the day, and was well aware that it wasn't hard to write a program that could play Breakout—but this was not what the DeepMind team had done. Instead, they'd created a blank-slate AI that knew nothing about this game—or about any other games, or even about *concepts* such as games, paddles, bricks or balls. All their AI knew was that a long list of numbers got fed into it at regular intervals: the current score and a long list of numbers which we (but not the AI) would recognize as specifications of how different parts of the screen were colored. The AI was simply told to maximize the score by outputting, at regular intervals, numbers which we (but not the AI) would recognize as codes for which keys to press.

Initially, the AI played terribly: it cluelessly jiggled the paddle back and forth seemingly at random and missed the ball almost every time. After a while, it seemed to be getting the idea that moving the paddle toward the ball was a good

idea, even though it still missed most of the time. But it kept improving with practice, and soon got better at the game than I'd ever been, infallibly returning the ball no matter how fast it approached. And then my jaw dropped: it figured out this amazing score-maximizing strategy of always aiming for the upper-left corner to drill a hole through the wall and let the ball get stuck bouncing between the back of the wall and the barrier behind it. This felt like a really intelligent thing to do. Indeed, Demis Hassabis later told me that the programmers on that DeepMind team didn't know this trick until they learned it from the AI they'd built. I recommend watching a video of this for yourself at the link I've provided.[1]

There was a human-like feature to this that I found somewhat unsettling: I was watching an AI that had a goal and learned to get ever better at achieving it, eventually outperforming its creators. In the previous chapter, we defined intelligence as simply the ability to accomplish complex goals, so in this sense, DeepMind's AI was growing more intelligent in front of my eyes (albeit merely in the very narrow sense of playing this particular game). In the first chapter, we encountered what computer scientists call *intelligent agents:* entities that collect information about their environment from sensors and then process this information to decide how to act back on their environment. Although DeepMind's game-playing AI lived in an extremely simple virtual world composed of bricks, paddles and balls, I couldn't deny that it was an intelligent agent.

DeepMind soon published their method and shared their code, explaining that it used a very simple yet powerful idea called *deep reinforcement learning*.[2] Basic reinforcement learning is a classic machine learning technique inspired by behaviorist psychology, where getting a positive reward increases your tendency to do something again and vice versa. Just like a dog learns to do tricks when this increases the likelihood of its getting encouragement or a snack from its owner soon, DeepMind's AI learned to move the paddle to catch the ball because this increased the likelihood of its getting more points soon. DeepMind combined this idea with deep learning: they trained a deep neural net, as in the previous chapter, to predict how many points would on average be gained by pressing each of the allowed keys on the keyboard, and then the AI selected whatever key the neural net rated as most promising given the current state of the game.

When I listed traits contributing to my own personal feeling of self-worth as a

human, I included the ability to tackle a broad range of unsolved problems. In contrast, being able to play Breakout and do nothing else constitutes extremely narrow intelligence. To me, the true importance of DeepMind's breakthrough is that deep reinforcement learning is a completely general technique. Sure enough, they let the exact same AI practice playing forty-nine different Atari games, and it learned to outplay their human testers on twenty-nine of them, from Pong to Boxing, Video Pinball and Space Invaders.

It didn't take long until the same AI idea had started proving itself on more modern games whose worlds were three-dimensional rather than two-dimensional. Soon DeepMind's San Francisco–based competitors at OpenAI released a platform called Universe, where DeepMind's AI and other intelligent agents can practice interacting with an entire computer as if it were a game: clicking on anything, typing anything, and opening and running whatever software they're able to navigate—firing up a web browser and messing around online, for example.

Looking to the future of deep reinforcement learning and improvements thereupon, there's no obvious end in sight. The potential isn't limited to virtual game worlds, since if you're a robot, life itself can be viewed as a game. Stuart Russell told me that his first major HS moment was watching the robot Big Dog run up a snow-covered forest slope, elegantly solving the legged locomotion problem that he himself had struggled to solve for many years.[3] Yet when that milestone was reached in 2008, it involved huge amounts of work by clever programmers. After DeepMind's breakthrough, there's no reason why a robot can't ultimately use some variant of deep reinforcement learning to teach itself to walk without help from human programmers: all that's needed is a system that gives it points whenever it makes progress. Robots in the real world similarly have the potential to learn to swim, fly, play ping-pong, fight and perform a nearly endless list of other motor tasks without help from human programmers. To speed things up and reduce the risk of getting stuck or damaging themselves during the learning process, they would probably do the first stages of their learning in virtual reality.

Intuition, Creativity and Strategy

Another defining moment for me was when the DeepMind AI system AlphaGo won a five-game Go match against Lee Sedol, generally considered the top player in the world in the early twenty-first century.

It was widely expected that human Go players would be dethroned by machines at some point, since it had happened to their chess-playing colleagues two decades earlier. However, most Go pundits predicted that it would take another decade, so AlphaGo's triumph was a pivotal moment for them as well as for me. Nick Bostrom and Ray Kurzweil have both emphasized how hard it can be to see AI breakthroughs coming, which is evident from interviews with Lee Sedol himself before and after losing the first three games:

- October 2015: "Based on its level seen…I think I will win the game by a near landslide."
- February 2016: "I have heard that Google DeepMind's AI is surprisingly strong and getting stronger, but I am confident that I can win at least this time."
- March 9, 2016: "I was very surprised because I didn't think I would lose."
- March 10, 2016: "I'm quite speechless…I am in shock. I can admit that…the third game is not going to be easy for me."
- March 12, 2016: "I kind of felt powerless."

Within a year after playing Lee Sedol, a further improved AlphaGo had played all twenty top players in the world without losing a single match.

Why was this such a big deal for me personally? Well, I confessed above that I view intuition and creativity as two of my core human traits, and as I'll now explain, I feel that AlphaGo displayed both.

Go players take turns placing black and white stones on a 19-by-19 board (see figure 3.2). There are vastly more possible Go positions than there are atoms in our Universe, which means that trying to analyze all interesting sequences of future moves rapidly gets hopeless. Players therefore rely heavily on

subconscious intuition to complement their conscious reasoning, with experts developing an almost uncanny feel for which positions are strong and which are weak. As we saw in the last chapter, the results of deep learning are sometimes reminiscent of intuition: a deep neural network might determine that an image portrays a cat without being able to explain why. The DeepMind team therefore gambled on the idea that deep learning might be able to recognize not merely cats, but also strong Go positions. The core idea that they built into AlphaGo was to marry the intuitive power of deep learning with the logical power of GOFAI—which stands for what's humorously known as "Good Old-Fashioned AI" from before the deep-learning revolution. They used a massive database of Go positions from both human play and games where AlphaGo had played a clone of itself, and trained a deep neural network to predict from each position the probability that white would ultimately win. They also trained a separate network to predict likely next moves. They then combined these networks with a GOFAI method that cleverly searched through a pruned list of likely future-move sequences to identify the next move that would lead to the strongest position down the road.

Figure 3.2: DeepMind's AlphaGo AI made a highly creative move on line 5, in defiance of millennia of human wisdom, which about fifty moves later proved crucial to its defeat of Go legend Lee Sedol.

This marriage of intuition and logic gave birth to moves that were not merely powerful, but in some cases also highly creative. For example, millennia of Go wisdom dictate that early in the game, it's best to play on the third or fourth line from an edge. There's a trade-off between the two: playing on the third line helps with short-term territory gain toward the side of the board, while playing on the fourth helps with long-term strategic influence toward the center.

In the thirty-seventh move of the second game, AlphaGo shocked the Go world by defying that ancient wisdom and playing on the fifth line (figure 3.2), as if it were even more confident than a human in its long-term planning abilities and therefore favored strategic advantage over short-term gain. Commentators were stunned, and Lee Sedol even got up and temporarily left the room.[4] Sure enough, about fifty moves later, fighting from the lower left-hand corner of the board ended up spilling over and connecting with that black stone from move thirty-seven! And that motif is what ultimately won the game, cementing the legacy of AlphaGo's fifth-row move as one of the most creative in Go history.

Because of its intuitive and creative aspects, Go is viewed more as an art form than just another game. It was considered one of the four “essential arts” in ancient China, together with painting, calligraphy and *qin* music, and it remains hugely popular in Asia, with almost 300 million people watching the first game between AlphaGo and Lee Sedol. As a result, the Go world was quite shaken by the outcome, and viewed AlphaGo’s victory as a profound milestone for humanity. Ke Jie, the world’s top-ranked Go player at the time, had this to say:[5] “Humanity has played Go for thousands of years, and yet, as AI has shown us, we have not yet even scratched the surface…The union of human and computer players will usher in a new era…Together, man and AI can find the truth of Go.” Such fruitful human-machine collaboration indeed appears promising in many areas, including science, where AI can hopefully help us humans deepen our understanding and realize our ultimate potential.

To me, AlphaGo also teaches us another important lesson for the near future: combining the intuition of deep learning with the logic of GOFAI can produce second-to-none *strategy*. Because Go is one of the ultimate strategy games, AI is now poised to graduate and challenge (or help) the best human strategists even beyond game boards—for example with investment strategy, political strategy and military strategy. Such real-world strategy problems are typically complicated by human psychology, missing information and factors that need to be modeled as random, but poker-playing AI systems have already demonstrated that none of these challenges are insurmountable.

Natural Language

Yet another area where AI progress has recently stunned me is language. I fell in love with travel early in life, and curiosity about other cultures and languages formed an important part of my identity. I was raised speaking Swedish and English, was taught German and Spanish in school, learned Portuguese and Romanian through two marriages and taught myself some Russian, French and Mandarin for fun.

But the AI has been reaching, and after an important discovery in 2016, there are almost no lazy languages that I can translate between better than the system of the AI developed by the equipment of the brain of Google.

Did I make myself crystal clear? I was actually trying to say this:

But AI has been catching up with me, and after a major breakthrough in 2016, there are almost no languages left that I can translate between better than the AI system developed by the Google Brain team.

However, I first translated it to Spanish and back using an app that I installed on my laptop a few years ago. In 2016, the Google Brain team upgraded their free Google Translate service to use deep recurrent neural networks, and the improvement over older GOFAI systems was dramatic:[6]

But AI has been catching up on me, and after a breakthrough in 2016, there are almost no languages left that can translate between better than the AI system developed by the Google Brain team.

As you can see, the pronoun “I” got lost during the Spanish detour, which unfortunately changed the meaning. Close, but no cigar! However, in defense of Google’s AI, I’m often criticized for writing unnecessarily long sentences that are hard to parse, and I picked one of my most confusingly convoluted ones for this example. For more typical sentences, their AI often translates impeccably. As a result, it created quite a stir when it came out, and it’s helpful enough to be used by hundreds of millions of people daily. Moreover, courtesy of recent progress in deep learning for speech-to-text and text-to-speech conversion, these users can now speak to their smartphones in one language and listen to the translated result.

Natural language processing is now one of the most rapidly advancing fields of AI, and I think that further success will have a large impact because language

is so central to being human. The better an AI gets at linguistic prediction, the better it can compose reasonable email responses or continue a spoken conversation. This might, at least to an outsider, give the appearance of human thought taking place. Deep-learning systems are thus taking baby steps toward passing the famous Turing test, where a machine has to converse well enough in writing to trick a person into thinking that it too is human.

Language-processing AI still has a long way to go, though. Although I must confess that I feel a bit deflated when I'm out-translated by an AI, I feel better once I remind myself that, so far, it doesn't *understand* what it's saying in any meaningful sense. From being trained on massive data sets, it discovers patterns and relations involving words without ever relating these words to anything in the real world. For example, it might represent each word by a list of a thousand numbers that specify how similar it is to certain other words. It may then conclude from this that the difference between "king" and "queen" is similar to that between "husband" and "wife"—but it still has no clue what it means to be male or female, or even that there is such a thing as a physical reality out there with space, time and matter.

Since the Turing test is fundamentally about deception, it has been criticized for testing human gullibility more than true artificial intelligence. In contrast, a rival test called the *Winograd Schema Challenge* goes straight for the jugular, homing in on that commonsense understanding that current deep-learning systems tend to lack. We humans routinely use real-world knowledge when parsing a sentence, to figure out what a pronoun refers to. For example, a typical Winograd challenge asks what "they" refers to here:

1. "The city councilmen refused the demonstrators a permit because they feared violence."
2. "The city councilmen refused the demonstrators a permit because they advocated violence."

There's an annual AI competition to answer such questions, and AIs still perform miserably.[7] This precise challenge, understanding what refers to what, torpedoed even GoogleTranslate when I replaced Spanish with Chinese in my example above:

But the AI has caught up with me, after a major break in 2016, with almost no language, I could translate the AI system than developed by the Google Brain

team.

Please try it yourself at https://translate.google.com now that you're reading the book and see if Google's AI has improved! There's a good chance that it has, since there are promising approaches out there for marrying deep recurrent neural nets with GOFAI to build a language-processing AI that includes a world model.

Opportunities and Challenges

These three examples were obviously just a sampler, since AI is progressing rapidly across many important fronts. Moreover, although I've mentioned only two companies in these examples, competing research groups at universities and other companies often weren't far behind. A loud sucking noise can be heard in computer science departments around the world as Apple, Baidu, DeepMind, Facebook, Google, Microsoft and others use lucrative offers to vacuum off students, postdocs and faculty.

It's important not to be misled by the examples I've given into viewing the history of AI as periods of stagnation punctuated by the occasional breakthrough. From my vantage point, I've instead been seeing fairly steady progress for a long time—which the media report as a breakthrough whenever it crosses the threshold of enabling a new imagination-grabbing application or useful product. I therefore consider it likely that brisk AI progress will continue for many years. Moreover, as we saw in the last chapter, there's no fundamental reason why this progress can't continue until AI matches human abilities on most tasks.

Which raises the question: How will this impact us? How will near-term AI progress change what it means to be human? We've seen that it's getting progressively harder to argue that AI completely lacks goals, breadth, intuition, creativity or language—traits that many feel are central to being human. This means that even in the near term, long before any AGI can match us at all tasks, AI might have a dramatic impact on how we view ourselves, on what we can do when complemented by AI and on what we can earn money doing when competing against AI. Will this impact be for the better or for the worse? What near-term opportunities and challenges will this present?

Everything we love about civilization is the product of human intelligence, so if we can amplify it with artificial intelligence, we obviously have the potential to make life even better. Even modest progress in AI might translate into major improvements in science and technology and corresponding reductions of accidents, disease, injustice, war, drudgery and poverty. But in order to reap these benefits of AI without creating new problems, we need to answer many important questions. For example:

1. How can we make future AI systems more robust than today's, so that they do what we want without crashing, malfunctioning or getting hacked?
2. How can we update our legal systems to be more fair and efficient and to keep pace with the rapidly changing digital landscape?
3. How can we make weapons smarter and less prone to killing innocent civilians without triggering an out-of-control arms race in lethal autonomous weapons?
4. How can we grow our prosperity through automation without leaving people lacking income or purpose?

Let's devote the rest of this chapter to exploring each of these questions in turn. These four near-term questions are aimed mainly at computer scientists, legal scholars, military strategists and economists, respectively. However, to help get the answers we need by the time we need them, everybody needs to join this conversation, because as we'll see, the challenges transcend all traditional boundaries—both between specialties and between nations.

Bugs vs. Robust AI

Information technology has already had great positive impact on virtually every sector of our human enterprise, from science to finance, manufacturing, transportation, healthcare, energy and communication, and this impact pales in comparison to the progress that AI has the potential to bring. But the more we come to rely on technology, the more important it becomes that it's robust and trustworthy, doing what we want it to do.

Throughout human history, we've relied on the same tried-and-true approach to keeping our technology beneficial: learning from mistakes. We invented fire, repeatedly messed up, and then invented the fire extinguisher, fire exit, fire alarm and fire department. We invented the automobile, repeatedly crashed, and then invented seat belts, air bags and self-driving cars. Up until now, our technologies have typically caused sufficiently few and limited accidents for their harm to be outweighed by their benefits. As we inexorably develop ever more powerful technology, however, we'll inevitably reach a point where even a single accident could be devastating enough to outweigh all benefits. Some argue that accidental global nuclear war would constitute such an example. Others argue that a bioengineered pandemic could qualify, and in the next chapter, we'll explore the controversy around whether future AI could cause human extinction. But we need not consider such extreme examples to reach a crucial conclusion: as technology grows more powerful, we should rely less on the trial-and-error approach to safety engineering. In other words, *we should become more proactive than reactive*, investing in safety research aimed at preventing accidents from happening even once. This is why society invests more in nuclear-reactor safety than mousetrap safety.

This is also the reason why, as we saw in chapter 1, there was strong community interest in AI-safety research at the Puerto Rico conference. Computers and AI systems have always crashed, but this time is different: AI is gradually entering the real world, and it's not merely a nuisance if it crashes the power grid, the stock market or a nuclear weapons system. In the rest of this section, I want to introduce you to the four main areas of technical AI-safety research that are dominating the current AI-safety discussion and that are being pursued around the world: *verification*, *validation*, *security* and *control*.[*1] To

prevent things from getting too nerdy and dry, let’s do this by exploring past successes and failures of information technology in different areas, as well as valuable lessons we can learn from them and research challenges that they pose.

Although most of these stories are old, involving low-tech computer systems that almost nobody would refer to as AI and that caused few, if any, casualties, we’ll see that they nonetheless teach us valuable lessons for designing safe and powerful future AI systems whose failures could be truly catastrophic.

AI for Space Exploration

Let's start with something close to my heart: space exploration. Computer technology has enabled us to fly people to the Moon and to send unmanned spacecraft to explore all the planets of our Solar System, even landing on Saturn's moon Titan and on a comet. As we'll explore in chapter 6, future AI may help us explore other solar systems and galaxies—if it's bug-free. On June 4, 1996, scientists hoping to research Earth's magnetosphere cheered jubilantly as an Ariane 5 rocket from the European Space Agency roared into the sky with the scientific instruments they'd built. Thirty-seven seconds later, their smiles vanished as the rocket exploded in a fireworks display costing hundreds of millions of dollars.[8] The cause was found to be buggy software manipulating a number that was too large to fit into the 16 bits allocated for it.[9] Two years later, NASA's Mars Climate Orbiter accidentally entered the Red Planet's atmosphere and disintegrated because two different parts of the software used different units for force, causing a 445% error in the rocket-engine thrust control.[10] This was NASA's second super-expensive bug: their Mariner 1 mission to Venus exploded after launch from Cape Canaveral on July 22, 1962, after the flight-control software was foiled by an incorrect punctuation mark.[11] As if to show that not only westerners had mastered the art of launching bugs into space, the Soviet Phobos 1 mission failed on September 2, 1988. This was the heaviest interplanetary spacecraft ever launched, with the spectacular goal of deploying a lander on Mars' moon Phobos—all thwarted when a missing hyphen caused the "end-of-mission" command to be sent to the spacecraft while it was en route to Mars, shutting down all of its systems.[12]

What we learn from these examples is the importance of what computer scientists call *verification:* ensuring that software fully satisfies all the expected requirements. The more lives and resources are at stake, the higher confidence we want that the software will work as intended. Fortunately, AI can help automate and improve the verification process. For example, a complete, general-purpose operating-system kernel called *seL4* has recently been mathematically checked against a formal specification to give a strong guarantee against crashes and unsafe operations: although it doesn't yet come with the bells and whistles of Microsoft Windows and Mac OS, you can rest assured that it won't give you what's affectionately known as "the blue screen of death" or

“the spinning wheel of doom.” The U.S. Defense Advanced Research Projects Agency (DARPA) has sponsored the development of a set of open-source high-assurance tools called HACMS (high-assurance cyber military systems) that are provably safe. An important challenge is to make such tools sufficiently powerful and easy to use that they’ll get widely deployed. Another challenge is that the very task of verification will itself get more difficult as software moves into robots and new environments, and as traditional preprogrammed software gets replaced by AI systems that keep learning, thereby changing their behavior, as in chapter 2.

AI for Finance

Finance is another area that's been transformed by information technology, allowing resources to be efficiently reallocated across the globe at the speed of light and enabling affordable financing for everything from mortgages to startup companies. Progress in AI is likely to offer great future profit opportunities from financial trading: most stock market buy/sell decisions are now made automatically by computers, and my graduating MIT students routinely get tempted by astronomical starting salaries to improve algorithmic trading.

Verification is important for financial software as well, which the American firm Knight Capital learned the hard way on August 1, 2012, by losing $440 million in forty-five minutes after deploying unverified trading software.[13] The trillion-dollar "Flash Crash" of May 6, 2010, was noteworthy for a different reason. Although it caused massive disruptions for about half an hour before markets stabilized, with shares of some prominent companies such Procter & Gamble swinging in price between a penny and $100,000,[14] the problem wasn't caused by bugs or computer malfunctions that verification could have avoided. Instead, it was caused by expectations being violated: automatic trading programs from many companies found themselves operating in an unexpected situation where their assumptions weren't valid—for example, the assumption that if a stock exchange computer reported that a stock had a price of one cent, then that stock really was worth one cent.

The flash crash illustrates the importance of what computer scientists call *validation:* whereas verification asks "Did I build the system right?," validation asks "Did I build the right system?"[*2] For example, does the system rely on assumptions that might not always be valid? If so, how can it be improved to better handle uncertainty?

AI for Manufacturing

Needless to say, AI holds great potential for improving manufacturing, by controlling robots that enhance both efficiency and precision. Ever-improving 3-D printers can now make prototypes of anything from office buildings to micromechanical devices smaller than a salt grain.[15] While huge industrial robots build cars and airplanes, affordable computer-controlled mills, lathes, cutters and the like are powering not merely factories, but also the grassroots "maker movement," where local enthusiasts materialize their ideas at over a thousand community-run "fab labs" around the world.[16] But the more robots we have around us, the more important it becomes that we verify and validate their software. The first person known to have been killed by a robot was Robert Williams, a worker at a Ford plant in Flat Rock, Michigan. In 1979, a robot that was supposed to retrieve parts from a storage area malfunctioned, and he climbed into the area to get the parts himself. The robot silently began operating and smashed his head, continuing for thirty minutes until his co-workers discovered what had happened.[17] The next robot victim was Kenji Urada, a maintenance engineer at a Kawasaki plant in Akashi, Japan. While working on a broken robot in 1981, he accidentally hit its on switch and was crushed to death by the robot's hydraulic arm.[18] In 2015, a twenty-two-year-old contractor at one of Volkswagen's production plants in Baunatal, Germany, was working on setting up a robot to grab auto parts and manipulate them. Something went wrong, causing the robot to grab him and crush him to death against a metal plate.[19]

Although these accidents are tragic, it's important to note that they make up a minuscule fraction of all industrial accidents. Moreover, industrial accidents have *decreased* rather than increased as technology has improved, dropping from about 14,000 deaths in 1970 to 4,821 in 2014 in the United States.[20] The three above-mentioned accidents show that adding intelligence to otherwise dumb machines should be able to further improve industrial safety, by having robots learn to be more careful around people. All three accidents could have been avoided with better validation: the robots caused harm not because of bugs or malice, but because they made invalid assumptions—that the person wasn't present or that the person was an auto part.

Figure 3.3: Whereas traditional industrial robots are expensive and hard to program, there's a trend toward cheaper AI-powered ones that can learn what to do from workers with no programming experience.

AI for Transportation

Although AI can save many lives in manufacturing, it can potentially save even more in transportation. Car accidents alone took over 1.2 million lives in 2015, and aircraft, train and boat accidents together killed thousands more. In the United States, with its high safety standards, motor vehicle accidents killed about 35,000 people last year—seven times more than all industrial accidents combined.[21] When we had a panel discussion about this in Austin, Texas, at the 2016 annual meeting of the Association for the Advancement of Artificial Intelligence, the Israeli computer scientist Moshe Vardi got quite emotional about it and argued that not only *could* AI reduce road fatalities, but it *must:* "It's a moral imperative!" he exclaimed. Because almost all car crashes are caused by human error, it's widely believed that AI-powered self-driving cars can eliminate at least 90% of road deaths, and this optimism is fueling great progress toward actually getting self-driving cars out on the roads. Elon Musk envisions that future self-driving cars will not only be safer, but will also earn money for their owners while they're not needed, by competing with Uber and Lyft.

So far, self-driving cars do indeed have a better safety record than human drivers, and the accidents that have occurred underscore the importance and difficulty of validation. The first fender bender caused by a Google self-driving car took place on February 14, 2016, because it made an incorrect assumption about a bus: that its driver would yield when the car pulled out in front of it. The first lethal crash caused by a self-driving Tesla, which rammed into the trailer of a truck crossing the highway on May 7, 2016, was caused by two bad assumptions:[22] that the bright white side of the trailer was merely part of the bright sky, and that the driver (who was allegedly watching a Harry Potter movie) was paying attention and would intervene if something went wrong.[*3]

But sometimes good verification and validation aren't enough to avoid accidents, because we also need good *control:* ability for a human operator to monitor the system and change its behavior if necessary. For such *human-in-the-loop* systems to work well, it's crucial that the human-machine communication be effective. In this spirit, a red light on your dashboard will conveniently alert you if you accidentally leave the trunk of your car open. In contrast, when the British car ferry *Herald of Free Enterprise* left the harbor of Zeebrugge on March 6, 1987, with her bow doors open, there was no warning light or other

visible warning for the captain, and the ferry capsized soon after leaving the harbor, killing 193 people.[23]

Another tragic control failure that might have been avoided by better machine-human communication occurred during the night of June 1, 2009, when Air France Flight 447 crashed into the Atlantic Ocean, killing all 228 on board. According to the official accident report, “the crew never understood that they were stalling and consequently never applied a recovery manoeuvre”—which would have involved pushing down the nose of the aircraft—until it was too late. Flight safety experts speculated that the crash might have been avoided had there been an “angle-of-attack” indicator in the cockpit, showing the pilots that the nose was pointed too far upward.[24]

When Air Inter Flight 148 crashed into the Vosges Mountains near Strasbourg in France on January 20, 1992, killing 87 people, the cause wasn’t lack of machine-human communication, but a confusing user interface. The pilots entered “33” on a keypad because they wanted to descend at an angle of 3.3 degrees, but the autopilot interpreted this as 3,300 feet per minute because it was in a different mode—and the display screen was too small to show the mode and allow the pilots to realize their mistake.

AI for Energy

Information technology has done wonders for power generation and distribution, with sophisticated algorithms balancing production and consumption across the world's electrical grids, and sophisticated control systems keeping power plants operating safely and efficiently. Future AI progress is likely to make the "smart grid" even smarter, to optimally adapt to changing supply and demand even down to the level of individual rooftop solar panels and home-battery systems. But on Thursday, August 14, 2003, it was lights-out for about 55 million people in the United States and Canada, many of whom remained powerless for days. Here, too, the primary cause was determined to be failed machine-human communications: a software bug prevented the alarm system in an Ohio control room from alerting operators to the need to redistribute power before a minor problem (overloaded transmission lines hitting unpruned foliage) cascaded out of control.[25]

The partial nuclear meltdown in a reactor on Three Mile Island in Pennsylvania on March 28, 1979, led to about a billion dollars in cleanup cost and a major backlash against nuclear power. The final accident report identified multiple contributing factors, including confusion caused by a poor user interface.[26] In particular, the warning light that the operators thought indicated whether a safety-critical valve was open or closed merely indicated whether a signal had been sent to close the valve—so the operators didn't realize that the valve had gotten stuck open.

These energy and transportation accidents teach us that as we put AI in charge of ever more physical systems, we need to put serious research efforts into not only making the machines work well on their own, but also into making machines collaborate effectively with their human controllers. As AI gets smarter, this will involve not merely building good user interfaces for information sharing, but also figuring out how to optimally allocate tasks within human-computer teams—for example, identifying situations where control should be transferred, and for applying human judgment efficiently to the highest-value decisions rather than distracting human controllers with a flood of unimportant information.

AI for Healthcare

AI has huge potential for improving healthcare. Digitization of medical records has already enabled doctors and patients to make faster and better decisions, and to get instant help from experts around the world in diagnosing digital images. Indeed, the best experts for performing such diagnosis may soon be AI systems, given the rapid progress in computer vision and deep learning. For example, a 2015 Dutch study showed that computer diagnosis of prostate cancer using magnetic resonance imaging (MRI) was as good as that of human radiologists,[27] and a 2016 Stanford study showed that AI could diagnose lung cancer using microscope images even better than human pathologists.[28] If machine learning can help reveal relationships between genes, diseases and treatment responses, it could revolutionize personalized medicine, make farm animals healthier and enable more resilient crops. Moreover, robots have the potential to become more accurate and reliable surgeons than humans, even without using advanced AI. A wide variety of robotic surgeries have been successfully performed in recent years, often allowing precision, miniaturization and smaller incisions that lead to decreased blood loss, less pain and shorter healing time.

Alas, there have been painful lessons about the importance of robust software also in the healthcare industry. For example, the Canadian-built Therac-25 radiation therapy machine was designed to treat cancer patients in two different modes: either with a low-power beam of electrons or with a high-power beam of megavolt X-rays that was kept on target by a special shield. Unfortunately, unverified buggy software occasionally caused technicians to deliver the megavolt beam when they thought they were administering the low-power beam, and without the shield, which ended up claiming the lives of several patients.[29] Many more patients died from radiation overdoses at the National Oncologic Institute in Panama, where radiotherapy equipment using radioactive cobalt-60 was programmed to excessive exposure times in 2000 and 2001 because of a confusing user interface that hadn’t been properly validated.[30] According to a recent report,[31] robotic surgery accidents were linked to 144 deaths and 1,391 injuries in the United States between 2000 and 2013, with common problems including not only hardware issues such as electrical arcing and burnt or broken pieces of instruments falling into the patient, but also software problems such as uncontrolled movements and spontaneous powering-off.

The good news is that the rest of almost two million robotic surgeries covered by the report went smoothly, and robots appear to be making surgery more rather than less safe. According to a U.S. government study, bad hospital care contributes to over 100,000 deaths per year in the United States alone,[32] so the moral imperative for developing better AI for medicine is arguably even stronger than that for self-driving cars.

AI for Communication

The communication industry is arguably the one where computers have had the greatest impact of all so far. After the introduction of computerized telephone switchboards in the fifties, the internet in the sixties, and the World Wide Web in 1989, billions of people now go online to communicate, shop, read news, watch movies or play games, accustomed to having the world's information just a click away—and often for free. The emerging *internet of things* promises improved efficiency, accuracy, convenience and economic benefit from bringing online everything from lamps, thermostats and freezers to biochip transponders on farm animals.

These spectacular successes in connecting the world have brought computer scientists a fourth challenge: they need to improve not only verification, validation and control, but also *security* against malicious software ("malware") and hacks. Whereas the aforementioned problems all resulted from unintentional mistakes, security is directed at *deliberate malfeasance*. The first malware to draw significant media attention was the so-called Morris worm, unleashed on November 2, 1988, which exploited bugs in the UNIX operating system. It was allegedly a misguided attempt to count how many computers were online, and although it infected and crashed about 10% of the 60,000 computers that made up the internet back then, this didn't stop its creator, Robert Morris, from eventually getting a tenured professorship in computer science at MIT.

Other malware exploits vulnerabilities not in software but in people. On May 5, 2000, as if to celebrate my birthday, people got emails with the subject line "ILOVEYOU" from acquaintances and colleagues, and those Microsoft Windows users who clicked on the attachment "LOVE-LETTER-FOR-YOU.txt.vbs" unwittingly launched a script that damaged their computer and re-sent the email to everyone in their address book. Created by two young programmers in the Philippines, this worm infected about 10% of the internet, just as the Morris worm had done, but because the internet was a lot bigger by then, it became one of the greatest infections of all time, afflicting over 50 million computers and causing over $5 billion in damages. As you're probably painfully aware, the internet remains infested with countless kinds of infectious malware, which security experts classify into worms, Trojans, viruses and other intimidating-sounding categories, and the damage they cause ranges from

displaying harmless prank messages to deleting your files, stealing your personal information, spying on you and hijacking your computer to send out spam.

Whereas malware targets whatever computer it can, *hackers* attack specific targets of interest—recent high-profile examples including Target, TJ Maxx, Sony Pictures, Ashley Madison, the Saudi oil company Aramco and the U.S. Democratic National Committee. Moreover, the loots appear to be getting ever more spectacular. Hackers stole 130 million credit card numbers and other account information from Heartland Payment Systems in 2008, and breached over a billion(!) Yahoo! email accounts in 2013.[33] A 2014 hack of the U.S. Government's Office of Personnel Management breached personnel records and job application information for over 21 million people, allegedly including employees with top security clearances and the fingerprints of undercover agents.

As a result, I roll my eyes whenever I read about some new system being allegedly 100% secure and unhackable. Yet "unhackable" is clearly what we need future AI systems to be before we put them in charge of, say, critical infrastructure or weapons systems, so the growing role of AI in society keeps raising the stakes for computer security. While some hacks exploit human gullibility or complex vulnerabilities in newly released software, others enable unauthorized login to remote computers by taking advantage of simple bugs that lingered unnoticed for an embarrassingly long time. The "Heartbleed" bug lasted from 2012 to 2014 in one of the most popular software libraries for secure communication between computers, and the "Bashdoor" bug was built into the very operating system of Unix computers from 1989 until 2014. This means that AI tools for improved verification and validation will improve security as well.

Unfortunately, better AI systems can also be used to find new vulnerabilities and perform more sophisticated hacks. Imagine, for example, that you one day get an unusually personalized "phishing" email attempting to persuade you to divulge personal information. It's sent from your friend's account by an AI who's hacked it and is impersonating her, imitating her writing style based on an analysis of her other sent emails, and including lots of personal information about you from other sources. Might you fall for this? What if the phishing email appears to come from your credit card company and is followed up by a phone call from a friendly human voice that you can't tell is AI-generated? In the ongoing computer-security arms race between offense and defense, there's so far little indication that defense is winning.

Human-Level Intelligence?

We've explored in this chapter how AI has the potential to greatly improve our lives in the near term, as long as we plan ahead and avoid various pitfalls. But what about the longer term? Will AI progress eventually stagnate due to insurmountable obstacles, or will AI researchers ultimately succeed in their original goal of building human-level artificial general intelligence? We saw in the previous chapter how the laws of physics allow suitable clumps of matter to remember, compute and learn, and how they don't prohibit such clumps from one day doing so with greater intelligence than the matter clumps in our heads. If/when we humans will succeed in building such superhuman AGI is much less clear. We saw in the first chapter that we simply don't know yet, since the world's leading AI experts are divided, most of them making estimates ranging from decades to centuries and some even guessing never. Forecasting is tough because, when you're exploring uncharted territory, you don't know how many mountains separate you from your destination. Typically you see only the closest one, and need to climb it before you can discover your next obstacle.

What's the soonest it could happen? Even if we knew the best possible way to build human-level AGI using today's computer hardware, which we don't, we'd still need to have enough of it to provide the raw computational power needed. So what's the computational power of a human brain measured in the bits and FLOPS from chapter 2?[*4] This is a delightfully tricky question, and the answer depends dramatically on how we ask it:

- Question 1: How many FLOPS are needed to simulate a brain?
- Question 2: How many FLOPS are needed for human intelligence?
- Question 3: How many FLOPS can a human brain perform?

There have been lots of papers published on question 1, and they typically give answers in the ballpark of a hundred petaFLOPS, i.e., 10^{17} FLOPS.[58] That's about the same computational power as the Sunway TaihuLight (figure 3.7), the world's fastest supercomputer in 2016, which cost about $300 million. Even if we knew how to use it to simulate the brain of a highly skilled worker, we would

only profit from having the simulation do this person's job if we could rent the TaihuLight for less than her hourly salary. We may need to pay even more, because many scientists believe that to accurately replicate the intelligence of a brain, we can't treat it as a mathematically simplified neural-network model from chapter 2. Perhaps we instead need to simulate it at the level of individual molecules or even subatomic particles, which would require dramatically more FLOPS.

The answer to question 3 is easier: I'm painfully bad at multiplying 19-digit numbers, and it would take me many minutes even if you let me borrow pencil and paper. That would clock me in below 0.01 FLOPS—a whopping 19 orders of magnitude below the answer to question 1! The reason for the huge discrepancy is that brains and supercomputers are optimized for extremely different tasks. We get a similar discrepancy between these questions:

How well can a tractor do the work of a Formula One race car?

How well can a Formula One car do the work of a tractor?

So which of these two questions about FLOPS are we trying to answer to forecast the future of AI? Neither! If we wanted to simulate a human brain, we'd care about question 1, but to build human-level AGI, what matters is instead the one in the middle: question 2. Nobody knows its answer yet, but it may well be significantly cheaper than simulating a brain if we either adapt the software to be better matched to today's computers or build more brain-like hardware (rapid progress is being made on so-called neuromorphic chips).

Hans Moravec estimated the answer by making an apples-to-apples comparison for a computation that both our brain and today's computers can do efficiently: certain low-level image-processing tasks that a human retina performs in the back of the eyeball before sending its results to the brain via the optic nerve.[59] He figured that replicating a retina's computations on a conventional computer requires about a billion FLOPS and that the whole brain does about ten thousand times more computation than a retina (based on comparing volumes and numbers of neurons), so that the computational capacity of the brain is around 10^{13} FLOPS—roughly the power of an optimized $1,000 computer in 2015!

Figure 3.7: Sunway TaihuLight, the world's fastest supercomputer in 2016, whose raw computational power arguably exceeds that of the human brain.

In summary, there's absolutely no guarantee that we'll manage to build human-level AGI in our lifetime—or ever. But there's also no watertight argument that we won't. There's no longer a strong argument that we lack enough hardware firepower or that it will be too expensive. We don't know how far we are from the finish line in terms of architectures, algorithms and software, but current progress is swift and the challenges are being tackled by a rapidly growing global community of talented AI researchers. In other words, we can't dismiss the possibility that AGI will eventually reach human levels and beyond. Let's therefore devote the next chapter to exploring this possibility and what it might lead to!

THE BOTTOM LINE:

- Near-term AI progress has the potential to greatly improve our lives in myriad ways, from making our personal lives, power grids and financial markets more efficient to saving lives with self-driving cars, surgical bots and AI diagnosis systems.
- When we allow real-world systems to be controlled by AI, it's crucial that we learn to make AI more robust, doing what we want it to do. This boils down to solving tough technical problems related to verification, validation, security and control.
- This need for improved robustness is particularly pressing for AI-controlled weapon systems, where the stakes can be huge.
- Many leading AI researchers and roboticists have called for an international treaty banning certain kinds of autonomous weapons, to avoid an out-of-control arms race that could end up making convenient assassination machines available to everybody with a full wallet and an axe to grind.
- AI can make our legal systems more fair and efficient if we can figure out how to make robojudges transparent and unbiased.
- Our laws need rapid updating to keep up with AI, which poses tough legal questions involving privacy, liability and regulation.
- Long before we need to worry about intelligent machines replacing us altogether, they may increasingly replace us on the job market.
- This need not be a bad thing, as long as society redistributes a fraction of the AI-created wealth to make everyone better off.
- Otherwise, many economists argue, inequality will greatly increase.
- With advance planning, a low-employment society should be able to flourish not only financially, with people getting their sense of purpose from activities other than jobs.
- Career advice for today's kids: Go into professions that machines are bad at—those involving people, unpredictability and creativity.
- There's a non-negligible possibility that AGI progress will proceed to human levels and beyond—we'll explore that in the next chapter!

---

*1 If you want a more detailed map of the AI-safety research landscape, there's an interactive one here, developed in a community effort spearheaded by FLI's Richard Mallah: https://futureoflife.org/landscape.

*2 More precisely, verification asks if a system meets its specifications, whereas validation asks if the correct specifications were chosen.

*3 Even including this crash in the statistics, Tesla's Autopilot was found to reduce crashes by 40% when turned on: http://tinyurl.com/teslasafety.

*4 Recall that FLOPS are floating-point operations per second, say, how many 19-digit numbers can be multiplied each second.