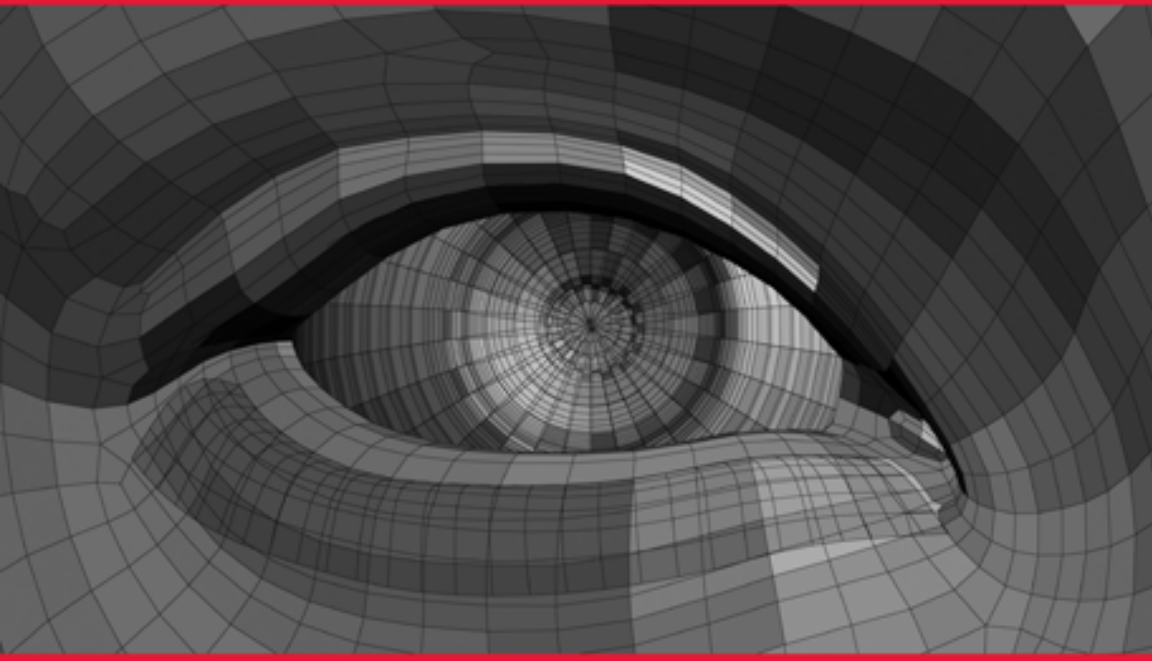
O'REILLY®

What Is Artificial Intelligence?



Mike Loukides & Ben Lorica

What Is Artificial Intelligence?

Mike Loukides and Ben Lorica

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

What Is Artificial Intelligence?
by Ben Lorica and Mike Loukides

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Nicole Tache
Production Editor: Melanie Yarbrough
Interior Designer: David Futato
Cover Designer: Randy Comer

June 2016: First Edition

Revision History for the First Edition
2016-06-15: First Release

978-1-491-96540-5

[LSI]

Table of Contents

What Is Artificial Intelligence? **1**
Capabilities and Limitations Today 1
Toward General Intelligence 2
To Train or Not to Train 4
The Meaning of Intelligence 6
Assistants or Actors? 7
Why the Surge of Interest? 9
Building Knowledge Databases 10
Producing Results 11
Ethics and Futures 13
Always in the Future 15

What Is Artificial Intelligence?

Defining artificial intelligence isn't just difficult; it's impossible, not the least because we don't really understand human intelligence. Paradoxically, advances in AI will help more to define what human intelligence isn't than what artificial intelligence is.

But whatever AI is, we've clearly made a lot of progress in the past few years, in areas ranging from computer vision to game playing. AI is making the transition from a research topic to the early stages of enterprise adoption. Companies such as Google and Facebook have placed huge bets on AI and are already using it in their products. But Google and Facebook are only the beginning: over the next decade, we'll see AI steadily creep into one product after another. We'll be communicating with bots, rather than scripted robo-dialers, and not realizing that they aren't human. We'll be relying on cars to plan routes and respond to road hazards. It's a good bet that in the next decades, some features of AI will be incorporated into every application that we touch and that we won't be able to do anything without touching an application.

Given that our future will inevitably be tied up with AI, it's imperative that we ask: Where are we now? What is the state of AI? And where are we heading?

Capabilities and Limitations Today

Descriptions of AI span several axes: strength (how intelligent is it?), breadth (does it solve a narrowly defined problem, or is it general?), training (how does it learn?), capabilities (what kinds of problems are we asking it to solve?), and autonomy (are AIs assistive technol-

ogies, or do they act on their own?). Each of these axes is a spectrum, and each point in this many-dimensional space represents a different way of understanding the goals and capabilities of an AI system.

On the strength axis, it's very easy to look at the results of the last 20 years and realize that we've made some extremely powerful programs. Deep Blue beat Garry Kasparov in chess; Watson beat the best Jeopardy champions of all time; AlphaGo beat Lee Sedol, arguably the world's best Go player. But all of these successes are limited. Deep Blue, Watson, and AlphaGo were all highly specialized, single-purpose machines that did one thing extremely well. Deep Blue and Watson can't play Go, and AlphaGo can't play chess or Jeopardy, even on a basic level. Their intelligence is very narrow, and can't be generalized. A lot of work has gone into using Watson for applications such as medical diagnosis, but it's still fundamentally a question-and-answer machine that must be tuned for a specific domain. Deep Blue has a lot of specialized knowledge about chess strategy and an encyclopedic knowledge of openings. AlphaGo was built with a more general architecture, but a lot of hand-crafted knowledge still made its way into the code. I don't mean to trivialize or undervalue their accomplishments, but it's important to realize what they haven't done.

We haven't yet created an artificial general intelligence that can solve a multiplicity of different kinds of problems. We still don't have a machine that can listen to recordings of humans for a year or two, and start speaking. While AlphaGo "learned" to play Go by analyzing thousands of games, and then playing thousands more against itself, the same software couldn't be used to master chess. The same general approach? Probably. But our best current efforts are far from a general intelligence that is flexible enough to learn without supervision, or flexible enough to choose what it wants to learn, whether that's playing board games or designing PC boards.

Toward General Intelligence

How do we get from narrow, domain-specific intelligence to more general intelligence? By "general intelligence," we don't necessarily mean human intelligence; but we do want machines that can solve different kinds of problems without being programmed with domain-specific knowledge. We want machines that can make

human judgments and decisions. That doesn't necessarily mean that AI systems will implement concepts like creativity, intuition, or instinct, which may have no digital analogs. A general intelligence would have the ability to follow multiple pursuits and to adapt to unexpected situations. And a general AI would undoubtedly implement concepts like "justice" and "fairness": we're already talking about the impact of AI on the legal system.

A self-driving car demonstrates the problems we're facing. To be self-driving, a car needs to integrate pattern recognition with other capabilities, including reasoning, planning, and memory. It needs to recognize patterns, so it can react to obstacles and street signs; it needs to reason, both to understand driving regulations and to solve problems like avoiding obstacles; it needs to plan a route from its current location to its destination, taking into account traffic and other patterns. It needs to do all of these repeatedly, updating its solutions constantly. However, even though a self-driving car incorporates just about all of AI, it doesn't have the flexibility we'd expect from a general intelligence system. You wouldn't expect a self-driving car to have a conversation or lay out your garden. Transfer learning, or taking results from one area and applying them to another, is very difficult. You could probably re-engineer many of the software components, but that only points out what's missing: our current AIs provide narrow solutions to specific problems; they aren't general problem solvers. You can add narrow AIs *ad infinitum* (a car could have a bot that talks about where to go; that makes restaurant recommendations; that plays chess with you so you don't get bored), but a pile of narrow intelligences will never add up to a general intelligence. General intelligence isn't about the number of abilities, but about integration between those abilities.

While approaches like neural networks were originally developed to mimic the human brain's processes, many AI initiatives have given up on the notion of imitating a biological brain. We don't know how brains work; neural networks are computationally useful, but they're not imitating human thought. In *Artificial Intelligence: A Modern Approach*, Peter Norvig and Stuart Russell write that "The quest for 'artificial flight' succeeded when the Wright brothers and others stopped imitating birds and started … learning about aerodynamics." Similarly, to make progress, AI need not focus on imitating the brain's biological processes, and instead try to understand the problems that the brain solves. It's a safe bet that humans use any number

of techniques to learn, regardless of what may be happening on the biological level. The same will probably be true of a general artificial intelligence: it will use pattern matching (like AlphaGo), it will use rule-based systems (like Watson), it will use exhaustive search trees (like Deep Blue). None of these techniques map directly onto human intelligence. What humans appear to do better than any computer is to build models of their world, and act on those models.

The next step past general intelligence is super-intelligence or hyper-intelligence. It's not clear how to distinguish super-intelligence from general intelligence. Would we expect a super-intelligence system to possess qualities like creativity and initiative? Given that we have trouble understanding human creativity, it's hard to think of machine creativity as a useful concept. Go experts described some of AlphaGo's moves as "creative"; however, they came out of exactly the same processes and patterns as all the other moves, not from looking at the game in a different way. Repeated application of the same algorithms can produce results that humans find surprising or unexpected, but merely being surprising isn't what we call "creativity."

It's easier to think of super-intelligence as a matter of scale. If we can create "general intelligence," it's easy to assume that it could quickly become thousands of times more powerful than human intelligence. Or, more precisely: either general intelligence will be significantly slower than human thought, and it will be difficult to speed it up either through hardware or software; or it will speed up quickly, through massive parallelism and hardware improvements. We'll go from thousand-core GPUs to trillions of cores on thousands of chips, with data streaming in from billions of sensors. In the first case, when speedups are slow, general intelligence might not be all that interesting (though it will have been a great ride for the researchers). In the second case, the ramp-up will be very steep and very fast.

To Train or Not to Train

AlphaGo's developers claimed to use a much more general approach to AI than Deep Blue: they produced a system that had minimal knowledge of Go strategy, but instead learned by observing Go games. That points toward the next big direction: can we get from supervised learning, where a machine is trained on labeled data, to

unsupervised learning, where a machine learns for itself how to group and structure data?

In a post on Facebook, Yann LeCun says, "We need to solve the unsupervised learning problem before we can even think of getting to true AI." To classify photos, an AI system is given millions of photos that have already been classified correctly; after learning from these classifications, it's given another set of tagged photos, to determine whether it can tag the test set correctly. What can a machine do without tagging? Can it discover what's important in a photo without metadata telling it "This is a bird, this is a plane, this is a flower"? Can a machine discover structure by observation with much less data, something that both humans and animals can do?

Both humans and animals can form models and abstractions from relatively little data: it doesn't take millions of images for us to recognize a new kind of bird, for example, or to find our way around a new city. Predicting future frames of a video, a problem researchers are now working on, would require an AI system to build an understanding of how the world works. Is it possible to develop a system that can respond to completely new situations, such as a car sliding unpredictably on ice? Is it possible to build a car that can drive without the benefit of a map? Humans can solve these problems, though they're not necessarily good at it. Unsupervised learning points to problems that can't just be solved by better, faster hardware, or by developers working with the current libraries.

There are approaches to learning that represent a point between supervised and unsupervised learning. In reinforcement learning, the system is given some value that represents a reward. Can a robot run across a field without falling? Can a car drive across town without a map? Rewards can be fed back into the system and used to maximize the probability of success. (OpenAI Gym is a promising framework for reinforcement learning.)

At one extreme, supervised learning means reproducing a set of tags, which is essentially pattern recognition, and prone to overfitting. At the other extreme, completely unsupervised learning means learning to reason inductively about a situation, and requires algorithmic breakthroughs. Semi-supervised learning (with minimal tags), or reinforcement learning (by sequential decision making) represent approaches between these extremes. We'll see how far they can take us.

The Meaning of Intelligence

What we mean by "intelligence" is a fundamental question. In a Radar post from 2014, Beau Cronin did an excellent job of summarizing the many definitions of AI. What we expect from artificial intelligence depends critically on what we want the AI to do. Discussions of AI almost always start with the Turing Test. Turing assumed that people would interact with a computer through a chat-like model: he assumed a conversation with the computer. This assumption places limitations on what we expect the computer to do: we don't expect it to drive cars or assemble circuits, for example. It's also an intentionally ambiguous test. The computer's answers might be evasive or just plain incorrect; being unerringly correct isn't the point. Human intelligences are also evasive and incorrect. We'd be unlikely to mistake an AI that was unerringly correct for a human.

If we assume that AI must be embodied in hardware that's capable of motion, such as a robot or an autonomous vehicle, we get a different set of criteria. We're asking the computer to perform a poorly defined task (like driving to the store) under its own control. We can already build AI systems that can do a better job of planning a route and driving than most humans. The one accident in which one of Google's autonomous vehicles was at fault occurred because the algorithms were modified to drive more like a human, and to take risks that the AI system would not normally have taken.

There are plenty of difficult driving problems that self-driving cars haven't solved: driving on a mountain road in a blizzard, for example. Whether the AI system is embodied in a car, a drone aircraft, or a humanoid robot, the problems it will face will be essentially similar: how to perform in safe, comfortable circumstances will be easy; how to perform in high-risk, dangerous situations will be much harder. Humans aren't good at those tasks, either; but while Turing would expect an AI in conversation to be evasive, or even answer questions incorrectly, vague or incorrect solutions while driving down a highway aren't acceptable.

AIs that can take physical action force us to think about robotic behavior. What sort of ethics govern autonomous robots? Asimov's laws of robotics? If we think a robot should never kill or injure a human, weaponized drones have already thrown that out the window. While the stereotypical question "if an accident is unavoidable, should an autonomous car crash into the baby or the grand-

mother?" is fake ethics, there are versions of the question that are more serious. Should a self-driving car plunge into a crowd to avoid an accident that might kill its passenger? It's easy to answer the question in the abstract, but it's hard to imagine humans buying a vehicle that will sacrifice them rather than injure bystanders. I doubt the robots will be expected to answer this question, but it will certainly be discussed in the board rooms of Ford, GM, Toyota, and Tesla.

We can define AI more simply by dispensing with the intricacies of conversational systems or autonomous robotic systems, and saying that AI is solely about building systems that answer questions and solve problems. Systems that can answer questions and reason about complex logic are the "expert systems" that we've been building for some years now, most recently embodied in Watson. (AlphaGo solves a different kind of problem.) However, as Beau Cronin points out, solving problems that humans find intellectually challenging is relatively easy; what's much more difficult is solving the problems that humans find easy. Few three year olds can play Go. All three year olds can recognize their parents—and without a substantial set of tagged images.

What we mean by "intelligence" depends strongly on what we want that intelligence to do. There is no single definition that's adequate for all of our goals. Without well-defined goals that tell us what we're trying to achieve, and let us measure whether we've achieved it, the transition from narrow AI to general AI is not going to be easy.

Assistants or Actors?

Press coverage of AI focuses on autonomous systems, machines that act on their own. With good reason: that's the fun, sexy, and somewhat scary face of AI. It's easy to watch AlphaGo, with a human servant to make its moves, and fantasize about a future dominated by machines. But there's something more to AI than autonomous devices that make humans obsolete. Where is the real value, artificial intelligence or intelligence augmentation? AI or IA? That question has been asked since the first attempts at AI and is explored in depth by John Markoff in *Machines of Loving Grace*. We may not want an AI system to make decisions; we may want to reserve decision making for ourselves. We may want AI that augments our intelligence by providing us with information, predicting the consequences of any course of action, and making recommenda-

tions, but leaving decisions to the humans. The Matrix notwithstanding, a future in which artificial intelligence is at our service, augmenting our intelligence rather than overruling it, is much more likely than a future in which we're the servants of an overreaching AI.

A GPS navigation system is an excellent example of an AI system that augments human intelligence. Given a good map, most humans can navigate from point A to point B, though our abilities leave a lot to be desired, particularly if we're in unfamiliar territory. Plotting the best route between two locations is a difficult problem, particularly when you account for problems like bad traffic and road conditions. But, with the exception of autonomous vehicles, we've never connected the navigation engine to the steering wheel. A GPS is strictly an assistive technology: it gives recommendations, not commands. Whenever you hear the GPS saying "recalculating route," a human has made a decision (or a mistake) that ignored the GPS recommendation, and the GPS is adapting.

Over the past few years, we've seen many applications that qualify as AI, in one sense or another. Almost anything that falls under the rubric of "machine learning" qualifies as artificial intelligence: indeed, "machine learning" was the name given to the more successful parts of AI back when the discipline fell into disrepute. You don't have to build something with a human voice, like Amazon's Alexa, to be AI. Amazon's recommendation engine is certainly AI. So is a web application like Stitchfix, which augments choices made by fashion experts with choices made by a recommendation engine. We've become accustomed to (and are frequently annoyed by) chat bots that handle customer service calls, more or less accurately. You'll probably end up talking to a human, but the secret is using the chat bot to get all the routine questions out of the way. There's no point in requiring a human to transcribe your address, your policy number, and other standard information: a computer can do it at least as accurately, if not more.

The next generation of assistants will be (and already is) semi-autonomous. Several years ago, Larry Page said the Star Trek computer was the ideal search engine: it was a computer that understood humans, had already digested all the information available, and could answer questions before they were even asked. If you have used Google Now, you were probably surprised the first time it told you to leave early for an appointment because the traffic was bad.

That requires looking across several different data sets: your current location, the location of your appointment (probably in your calendar or in your contacts list), Google's mapping data, current traffic conditions, and even chronological data about expected traffic patterns. The goal isn't answering a question; it's providing assistance before users even know they need it.

Why the Surge of Interest?

Why is AI currently such a hot topic, after having being in disrepute for a few decades of "AI winter"? Of course, AI was in the news briefly after Deep Blue, and again after Watson; but these fads didn't last. It's tempting to see the current rise of AI as another fad. That would ignore the changes of the past decade.

The rise of AI has depended on tremendous advances in computer hardware. It's tedious to recite the huge advances in performance and storage technology in the 30+ years since the start of the AI winter (which Wikipedia traces to 1984). But that's an unavoidable part of the story, particularly if you've seen the racks of machines that made up IBM's Watson. AlphaGo reportedly ran on 1,920 CPUs and 280 GPUs; the machine that beat Lee Sedol may have been even larger, and used custom hardware Google has developed for building neural networks. Even if AI algorithms are too slow to be productive on a typical laptop, it's easy and relatively inexpensive to allocate some serious computing horsepower on cloud platforms like AWS, GCE, and Azure. And machine learning was enabled, in part, by the ability to store vast amounts of data. In 1985, gigabytes were rare, and weighed hundreds of pounds; now gigabytes are commonplace, inexpensive, and tiny.

In addition to the ability to store and process data, we now have the ability to generate data. In the 80s, most photography was analog. Now it's all digital, and a lot of it is stored online, in services like Flickr, Google Photos, Apple Photos, Facebook, and more. Many online photos are already tagged with some descriptive text, making them a great dataset for training AI systems. Many of our conversations are also online, through Facebook, Twitter, and many chat services. As are our shopping histories. So we (or more precisely, Google, Apple, Yahoo, Facebook, Amazon, and others) have the data needed to train AI systems.

We've also made significant advances in algorithms. Neural networks aren't particularly new, but "deep learning" stacks up a series of networks, with feedback so the network automatically trains itself. Deep learning thus tries to solve one of the hardest human problems in machine learning: learning optimal representations and features from data. Processing a lot of data is easy, but feature learning is more of an art than a science. Deep learning automates some of that art.

Not only have we made progress in algorithms, the algorithms are implemented in widely available libraries, such as Caffe, TensorFlow, Theano, Scikit-Learn, MXNet, CNTK, and others. AI isn't limited to CS researchers in academic settings; increasingly, anyone can take part, as Pete Warden has shown. You don't need to know how to implement a complex algorithm and make it run reasonably well on your hardware. You just need to know how to install a library and tag training data. Just as the PC revolution itself took place when computers moved out of machine rooms and became accessible to the general public, the same process of democratization is producing a revolution in AI. As people from many backgrounds and environments experiment with AI, we'll see new kinds of applications. Some will seem like science fiction (though self-driving cars seemed like science fiction only a few years ago); there will certainly be new applications that we can't even imagine.

Building Knowledge Databases

The world is full of "dark data": unstructured information that doesn't live in nice, orderly databases. It's on websites, buried in tables, enshrined in photographs and movies; but it's not easily accessible to machine intelligence, or to any other intelligence. Projects like diffbot and deepdive use semi-supervised learning to find the structure in unstructured data—whether masses of scientific papers or the scrapings of many websites. Once they've created a database, that database can be accessed by more-conventional tools, whether APIs, SQL statements, or desktop applications.

Knowledge databases and graphs are already in use in many intelligent applications, including Google's Knowledge Graph. As we move toward conversational applications, the ability to unearth dark data and find structure in it will become even more critical. Using dark data effectively will be the key to moving from scripted and

narrowly purposed chat applications to applications that can take an arbitrary question and return an answer to the user. We might not see such an application as "understanding" the question, but applications like this will be at the heart of future assistive technologies. And they will rely on knowledge databases that have been scraped and structured by machine: the sheer volume of data involved will be beyond humans' tagging abilities.

Producing Results

Unlike the dark times of the AI winter, when data was limited and computers were slow, we're seeing successful AI systems everywhere. Google Translate is nowhere near as good as a human translator, but it frequently gives you a usable translation. While it hasn't been on the radar anywhere near as much, speech recognition systems are also commonplace, and surprisingly accurate; a year ago, Google claimed that an Android phone could correctly understand 92% of the questions it was asked. Given that a computer can correctly turn a question into text, the next step is to turn that question into an answer.

Similarly, image recognition and image processing have become commonplace. Despite some highly publicized and embarrassing mistakes, computer vision systems can identify faces with an accuracy that was unthinkable a few years ago. Granted, constraining the problem properly plays a huge role in this success: Facebook can identify faces in your photos because it's assuming that the people in your photos are likely to be your friends. Computer vision is (or will be) central to many applications of AI, from the mundane to the scary. Vision is obviously critical to autonomous vehicles; it's also critical to surveillance, auto-targeting drones, and other uncomfortable applications.

Deep learning and neural networks have attracted a lot of attention in the past year: they have enabled progress in computer vision, natural language, and other fields. However, almost anything that falls under the rubric of machine learning is artificial intelligence: classification and clustering algorithms, various kinds of decision trees, genetic algorithms, support vector machines, hierarchical temporal memory, and many others. These techniques can be used by themselves, or in combination with others. IBM's Watson is a good example of ensemble learning: it is a rule-based system that incorporates

many other algorithms, depending on the problem it is solving. The rules are largely hand-crafted, and the other algorithms need to be painstakingly tuned to get good results.

Impressive as Watson is, systems that require huge amounts of hand tuning are at best a stepping stone toward intelligence. Any general AI, and most narrow AIs, will probably combine many algorithms, rather than using a single, yet-to-be-discovered master algorithm. But the tuning required to get good results is a major limitation: Demis Hassabis, leader of the AlphaGo team, says that tuning is "almost like an art form." Is it really "artificial intelligence" if getting good results requires years of work, and only a few specialists (Hassabis says a few hundred) are capable of doing that work? The creation of an engine like Watson is science, but it also requires a lot of art. In addition, the need for manual optimization suggests that AIs built this way are inherently narrow, designed to solve a single problem. It's very difficult to imagine optimizing a "general intelligence" engine that can work on any problem. If you're optimizing, you're almost certainly optimizing for something, for some specific application.

Do advances in AI depend on better algorithms, or better hardware? The answer to that question is "both," if the question is even meaningful. Even though clock speeds have stalled, our ability to put more and more on a chip hasn't stalled: AlphaGo's 280 GPUs could easily mean 200,000 cores. More important, though, we've seen a lot of improvement in mathematical libraries and tools for using GPUs. We may also see the use of ASICs and FPGAs (application-specific integrated circuit and field-programmable gate arrays) in future AI engines. In turn, ASICs and FPGAs will be critical to embedding AI in hardware systems, many of which (think autonomous vehicles) will need to run in hard real-time.

But even if the hardware is better, we will still need algorithms that can be distributed across thousands or millions of nodes; we will need algorithms that can reprogram FPGAs on the fly, to adapt the hardware to the problems they are solving. MapReduce became popular for data analysis because it suggested a way to parallelize a large class of problems. Parallelism obviously works for AI; but what are its limits? The hard fact of parallelism is that the part of the program that can't be parallelized kills you. And the hallmark of most parallel algorithms is that you need a phase that collects partial results and generates a single result. AlphaGo may look at thousands

of alternatives when computing its next move; but at some point, it needs to look at all the alternatives, evaluate which is best, and present a single result. AlphaGo can take advantage of 280 GPUs; what about a machine with 280,000? After all, the largest AI systems we've built so far are a small fraction of the size of a rat brain, let alone a human brain. What about algorithms that don't lend themselves to parallelism as well as neural networks? How do you apply feedback in systems where each element of the pipeline is taking a different approach to the problem? Questions like these are likely to drive AI research in the near future.

Throwing more (and faster) hardware at AI algorithms is likely to get us better Go, chess, and Jeopardy players. We'll be able to classify images better and faster. But that's an improvement in the problems we can currently solve. Will more computing power get us from supervised to unsupervised learning? Will it lead us from narrow intelligence to general intelligence? That remains to be seen. Unsupervised learning is a hard problem, and it's not clear that it can be solved just by throwing more hardware at it. We're still looking for a "master algorithm" that may not exist.

Ethics and Futures

It is easy to get scared by talk of superhuman intelligence. And, according to some, it is time to decide what we want our machines to do, before it's too late. While this position may be oversimplified, it is very difficult to think about limiting devices that we can't build, and whose capabilities we can't imagine now, and may never understand in the future. It is also difficult to say "no," because I'm not aware of any technology that hasn't been invented because people thought better of it beforehand. At different times in history, people were afraid of many technologies that are now commonplace: at one point, many thought that traveling over 60 miles per hour would be fatal. Socrates was opposed to writing because he thought it would lead to forgetfulness: imagine what he would have thought of our technology!

But we can think about the future of AI, and how we will develop machines that assist us. Here are a few suggestions:

Most fears of a super-intelligent AI aren't really fears of a machine we neither know or understand; they are fears about human nature at its worst, coupled with unlimited power. We don't imagine a

machine that thinks thoughts we can't comprehend; we imagine an unbeatable Hitler or Stalin, whose thoughts we do comprehend. Our fears are essentially human fears: fears of omnipotent machines acting like humans.

That isn't to denigrate our fears, because we've seen that machine learning does learn from us. Microsoft's unfortunate Tay was a too-perfect example of a conversational AI bot that "learned" racism and bigotry from the people it talked to online. Google's image classification that identified a black couple as "gorillas" was the result of poor testing and a training set that didn't have enough properly tagged pictures of black people. Machines learn to be racist much the same way that humans do: because we teach them to be that way, whether intentionally or accidentally. That's a human problem, and it's one that's solvable. We can be more careful about what, and how, our artificial intelligences learn. We can be more careful about what's in our training sets, how those sets are tagged; and we can filter what kinds of answers we consider acceptable. None of this is terribly difficult; it just has to be done. What's more difficult in the current climate is reaching a consensus that racism and hatred are not OK.

That's a problem of human values, not machine intelligence. We can build machines that reflect our values: we do that already. Are they the values that we want to reflect? The White House's Report on Data Science, Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights, concludes with a section on the need for research on how to audit algorithms to "ensure that people are being treated fairly." As we move from "big data" to AI, the need to audit our algorithms and make sure they reflect values we support will only increase.

It is extremely important that research into artificial intelligence be open and visible to the public. Not because we believe the public will be less "afraid" of research that happens in the open (that may or may not be true), or because the public will somehow become "used to" the idea of super-intelligence; but because there is greater concern about research that goes on behind closed doors versus research that goes on in public. Indeed, Unethical Research suggests that the best way to create a healthy AI ecosystem is to publish ideas for creating malevolent machines. Research will continue to go on behind closed doors; it would be naive to think that military research and intelligence services aren't working on artificial intelligence. But we will be at the mercy of that research if there isn't also

AI research happening in the public sphere. (Whether an organization such as Google or Facebook constitutes "behind closed doors" or "in public view" is a debate worth having.) That's the point of OpenAI: "to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return." OpenAI is a dramatic and surprising answer to the fear of AI: push AI research forward as far as possible, but in public, make sure that the public sphere remains ahead of closed-door AI.

It is also important for research to be open and in the public sphere because research's origins often determine its application. Nuclear energy is a good example. It's possible to build safe, efficient nuclear reactors that are fueled by Thorium. But Thorium reactors were never built because they don't help you build bombs, and research into nuclear power was under the control of the defense department. A reactor that doesn't generate plutonium in usable quantities? Why would anyone want that? Again, it's naive to think that military and national intelligence services won't do excellent AI research. But if AI research becomes the sole province of the military, we will have excellent auto-targeting drones; if AI research becomes the sole province of national intelligence, we will have excellent systems for surreptitiously listening to and understanding conversations. Our imaginations will be limited about what else AI might do for us, and we will have trouble imagining AI applications other than murderous drones and the watchful ear of Big Brother. We may never develop intelligent medical systems or robotic nurses' aides.

If we want AI to serve humanity, it must be developed in public: as part of the larger community of AI researchers, and as part of a wider public discourse about our goals and aims. We must be careful not to build our own worst nightmare; but we need to realize that the nightmare is really just a more powerful version of ourselves.

Always in the Future

Mark Zuckerberg recently said that AI will be better than humans at most basic tasks in 5 to 10 years. He may be correct, but it's also clear that he's talking about narrow intelligence: specific tasks like speech recognition, image classification, and of course, game playing. He continues to say "That doesn't mean that the computers will be thinking..." Depending on who you talk to, a real general intelli-

gence is 10 to 50 years out. Given the difficulty of predicting the future of technology, the best answer is "more than 10 years," and possibly much more. When will human-level, machine intelligence (HLMI) be achieved? A recent survey of experts suggests that HLMI will occur (with 50% probability) sometime between 2040-2050." As Yann LeCun says, "Human-level general AI is several decades away."

So, when will we get there, if ever? A few years ago, Jason Huggins (@hugs) made a prescient remark about robots. Robots, he said, are always in the future. From time to time, bits of robotics break off and become part of the present; but when that happens, they're no longer considered robotics. In the 1920s, we would have considered a modern dishwasher a super-intelligent robot; now, it's just a dishwasher.

The same will inevitably happen for artificial intelligence. Indeed, it is already happening. I have avoided making a distinction between machine learning and AI; "machine learning" is a term that was applied to ideas from AI research back when AI was disreputable. Now, many of those ideas are commonplace. We don't think twice about Amazon recommendations or GPS navigation systems; we take those for granted. We may find Facebook and Google's ability to tag photos creepy, but we don't think AI when we see it. All serious chess players practice against chess programs; so do beginning Go players, and after AlphaGo's success, practicing against a computer will certainly extend to experts. These are artificial intelligences that have already broken off and become parts of our world. In doing so, AI morphed into IA (intelligence augmentation): autonomous technology that trumped human abilities became assistive.

Will we ever be able to point to something and say "Yes, that's artificial intelligence"? Yes, certainly; we can do that now. What's more important is that we will inevitably be surrounded by AI, bathed in it, even before we know it. We take plumbing for granted; we take electricity for granted; our children take streaming music for granted. We will take AI for granted, even as it becomes a larger part of our lives.

About the Authors

Mike Loukides is vice president of content strategy for O'Reilly Media, Inc. He has edited many highly regarded books on technical subjects that don't involve Windows programming. He's particularly interested in programming languages, Unix and what passes for Unix these days, and system and network administration. Mike is the author of *System Performance Tuning* and a coauthor of *Unix Power Tools*. Most recently, he's been fooling around with data and data analysis, languages like R, Mathematica, and Octave, and thinking about how to make books social.

Ben Lorica is the Chief Data Scientist of O'Reilly Media, and Program Director of Strata+Hadoop World and the O'Reilly Artificial Intelligence conference. He has applied business intelligence, data mining, machine learning, and statistical analysis in a variety of settings, including direct marketing, consumer and market research, targeted advertising, text mining, and financial engineering.