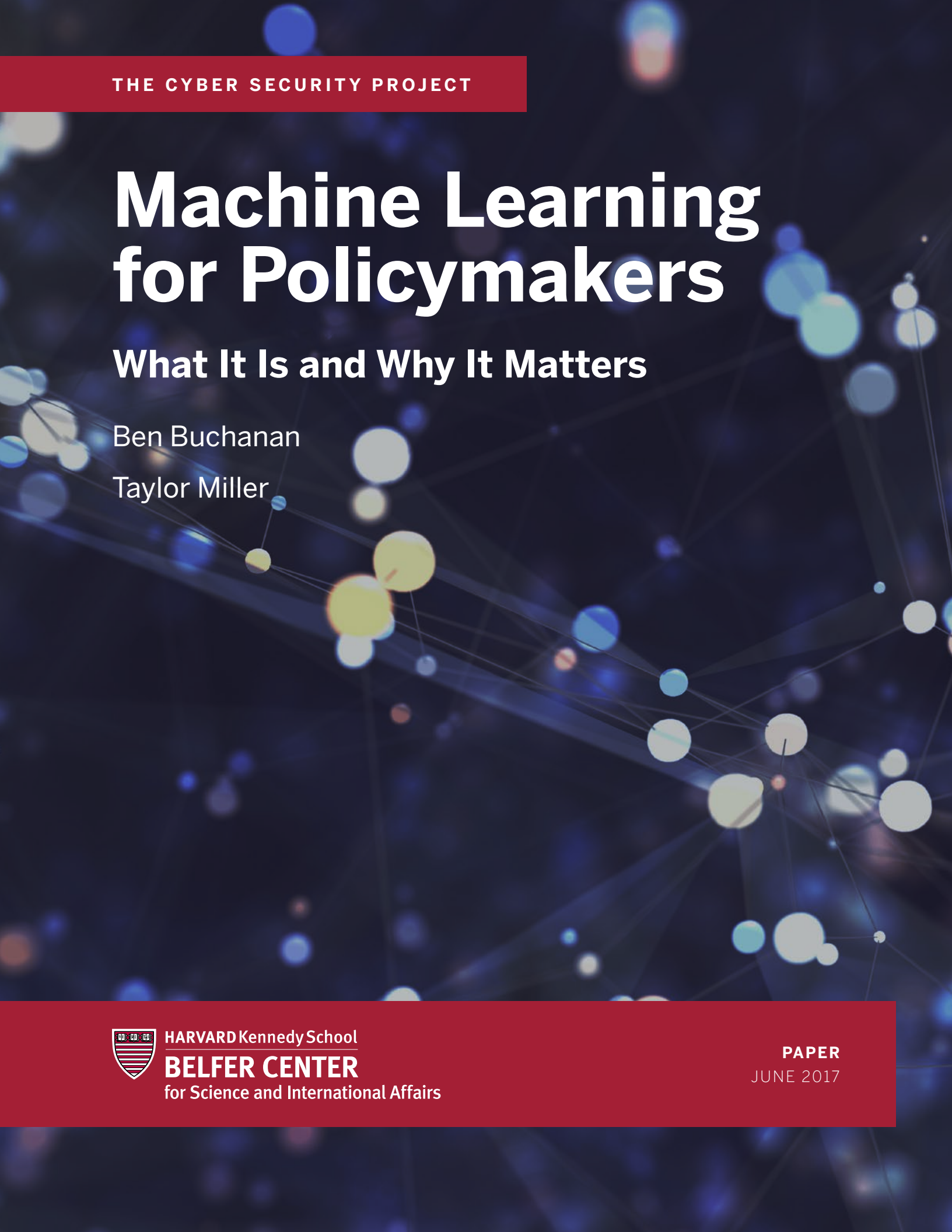


THE CYBER SECURITY PROJECT

Machine Learning for Policymakers

What It Is and Why It Matters

Ben Buchanan

Taylor Miller

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
JUNE 2017

The Cyber Security Project
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org/Cyber

Design & Layout by Andrew Facini

THE CYBER SECURITY PROJECT

Machine Learning for Policymakers

What It Is and Why It Matters

Ben Buchanan

Taylor Miller

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
JUNE 2017

About the Author

Ben Buchanan is a Postdoctoral Fellow at Harvard University's Cybersecurity Project at the Belfer Center for Science and International Affairs, where he conducts research on the intersection of cybersecurity and statecraft. His first book, The Cybersecurity Dilemma, was published by Oxford University Press in 2017. Previously, he has written on attributing cyber attacks, deterrence in cyber operations, cryptography, election cybersecurity, and the spread of malicious code between nations and non-state actors. He received his PhD in War Studies from King's College London, where he was a Marshall Scholar, and earned masters and undergraduate degrees from Georgetown University.

Taylor Miller is a second-year medical student at the Icahn School of Medicine at Mount Sinai. Before medical school, he worked on analyzing federal health policy in Washington, D.C., especially in the areas of healthcare access, insurance networks, health technology, and rural health. He currently conducts research on using machine learning to improve predictive modeling for patient outcomes in value-based care settings.

Acknowledgements

We would like to thank Tarun Chhabra, Teddy Collins, Bruce Schneier, Michael Sulmeyer, and Gabriella Roncone for their comments on an earlier draft of this paper. All errors remain ours alone.

Our thanks as well to GLG for arranging research interviews.

Table of Contents

Executive Summary

Machine learning matters. If nothing else, the drumbeat of headlines in recent years offers proof of this. In fields as diverse as healthcare, transportation, policing, and warfighting, machine learning algorithms have already had a significant impact. They seem poised to do more, and the particulars and the implications of this change deserve attention.

But machine learning can seem incomprehensible. As a type of artificial intelligence, it can be technical and obtuse. As a fast-changing discipline, it can appear to lack conceptual constants. As a domain of sometimes-inscrutable algorithms, it often hides the answer to one of its most pressing questions: why do machines do what they do?

While all of these are real concerns, we believe that not only is it possible for generalists to gain insight into machine learning, it is vital. This paper aims to enable that understanding. First, we introduce and differentiate three types of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. We show how each is well-suited to particular tasks and how the combination of different algorithms and architectures can lead to powerful results. Second, we examine how machine learning has already affected a disparate array of fields. We use these examples of success to introduce important concepts, such as deep learning, computer vision, and the importance of data.

With key concepts identified, we next examine how machine learning is poised to be of still greater significance in areas of importance to policymakers. The third section therefore considers the impact machine learning could make in warfighting, healthcare, and policing. New technologies will deeply impact how business is done, changing the nature of jobs and potentially improving overall outcomes. But in each area, there are specific policy, ethical, and technical challenges that must be addressed in order to achieve the best results. For example, the ethics of artificial intelligence in conflict, the challenges of data interoperability in healthcare, and the danger of bias in policing all

deserve attention. It is vital that policymakers have an understanding of the key facts of machine learning as they work through these sector-specific challenges.

Fourth, we outline some general matters that deserve attention when it comes to machine learning, such as bias, privacy, explainability, and security. In each of these areas, there are crucial questions and challenges. For example, there is tension between the improved accuracy that comes from taking all data into account and the increased unfairness from relying on data correlated with race to make predictions; a similar tension exists between increasing algorithms' accuracy and usefulness and protecting the privacy of individuals. All of this is complicated by the difficulty in understanding how machine learning algorithms come to certain conclusions—even when those conclusions are correct. Security is also an ever-growing issue. When it comes to managing these and other challenges, engaging with the discipline's foundational concepts is vital.

Fifth, we make recommendations to chart a path forward. It is essential that the bias in already-deployed machine learning algorithms be understood, and that ethics and impacts of machine learning are considered going forward. It is likewise essential that governments encourage the sharing of useful data, and look to how they can better deploy machine learning to improve their own operations. Finally, governments should encourage research and education in machine learning algorithms and applications, particularly those that enhance privacy, security, and explainability.

Few areas of national policymaking will remain untouched by artificial intelligence. Though the challenges it poses are complex, the opportunities it offers are tremendous. Simply put, machine learning is too important to ignore.

Introduction

Machine learning can spot cancer. It can translate complex texts. Drive cars. Beat the best human in the world at one of the most complex games ever invented. Devise alien-like designs to create more efficient physical structures. Save energy.

The science fiction writer and futurist Arthur C. Clarke wrote, "Any sufficiently advanced technology is indistinguishable from magic."[1] The accomplishments above can indeed at times seem magical, but they are not. These successes are the result of a combination of innovative algorithms, powerful computers, and rich data. This mixture of algorithms, computers, and data can also, when misapplied or when misconfigured, make significant mistakes with catastrophic consequences. To see machine learning as sorcery rather than as a powerful tool that must be wielded carefully and thoughtfully is to invite enormous risk.

Machine learning can also seem magical in another way: it can appear impossible to grasp. Our foundational premise in this paper is that this idea is false and dangerous. For each of the aforementioned achievements, and for several others, we will outline the concepts at play in a way that is accessible to generalists. Not only do we believe it is possible for non-specialists to gain intuition about how machine learning works, we think it is urgent. Several important principles are fundamental to understanding the power of machine learning, the opportunities it offers, and the new policy issues it raises.

We proceed as follows. The first section outlines the basics of how machine learning works. It provides some important background on artificial intelligence, and discusses three main types of machine learning algorithms. The second section considers the current state of affairs, identifying areas of great progress in machine learning and in the process distilling important foundational concepts. In so doing, we show the ways in which machine learning has already had an impact on a variety of challenges. Next, we turn to the future. The third section examines how machine learning will

1 Arthur C. Clarke, *Profiles of the Future* (New York: Macmillan, 1973).

affect areas of great importance to policymakers. In particular, we focus on warfighting, healthcare, and policing. The discussion of these three areas shows the breadth of the change still to come, and the need for policymaker engagement.

While there are sector-specific reasons for policymaker engagement on machine learning, there are also overarching ones. The fourth section examines the challenges that come with the technology. In particular, it articulates concerns about data availability, privacy, fairness, security, and economic impact that must be carefully managed. Each of these areas, if not addressed by technologists and policymakers, represents a way in which poorly designed or applied machine learning tools could cause real harm. We believe they all deserve significant attention. As such, our conclusion provides recommendations on how policymakers can begin to approach machine learning to best maximize its potential and overcome its dangers.

How Does Machine Learning Work?

Machine learning is the process of instructing computers to learn.[2] It exists at the intersection of computer science, statistics, and linear algebra, with insights from neuroscience and other fields as well. But unlike traditional software development, machine learning involves programming computers to teach themselves from data rather than instructing them to perform certain tasks in certain ways. Machine learning is traditionally focused on prediction and creating structure out of unstructured data.

In early efforts at artificial intelligence, researchers would attempt to instruct computers to act based on clear preset rules with fixed conditions. For example, a very basic spam filter program might be told to mark as spam every email with a subject that contains the full phrase "cheap imported drugs." This approach has the advantage of being straightforward, but it is also inflexible; spammers who instead use the subject "discounted imported drugs" can defeat the system. In contrast, a machine learning program forgoes these few predetermined rules, which are often too broad or too narrow to be effective. The modern machine learning program instead identifies on its own a large number of more subtle patterns and features in data given to it for training purposes. It then uses these insights to assess new data. When making an assessment, such as whether or not an email is spam, the machine learning program will evaluate all of the features of the new data and compare them to the patterns it has seen before.

For engineers, building a machine learning capability can often take a great deal of fine-tuning and experimentation, as well as the use of conceptually interesting techniques (some of which will be discussed below). In addition, many machine learning techniques require computing power that has only very recently become available to those outside of government, even though the concepts themselves are older. To simplify the enormous amount of complexity and variation in these systems, machine learning algorithms are often divided into three broad categories: *supervised*

2 This section recapitulates the basics of machine learning. For more detailed overviews, see Tom Michael Mitchell, *The Discipline of Machine Learning*, vol. 9 (Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006). Richard S Sutton and Andrew G Barto, *Reinforcement Learning: An Introduction*, vol. 1 (MIT press Cambridge, 1998).

learning, *unsupervised learning*, and *reinforcement learning*. Each of these is a different method of applying machines to a problem.

Supervised Learning

Supervised learning algorithms enable machines to make predictions or assessments; they are widely used in everyday life, from voice recognition to email spam filters to medical predictions. The "supervised" part of the name comes from the fact that each piece of data given to the algorithm also contains the correct answer about the characteristic of interest, such as whether an email is spam or not, so that the algorithm can learn from past data and test itself by making predictions. To do this, the computer is usually given three things:

- A set of data to learn from. The data can be provided in a table or spreadsheet format, but must be labeled with the correct categories.

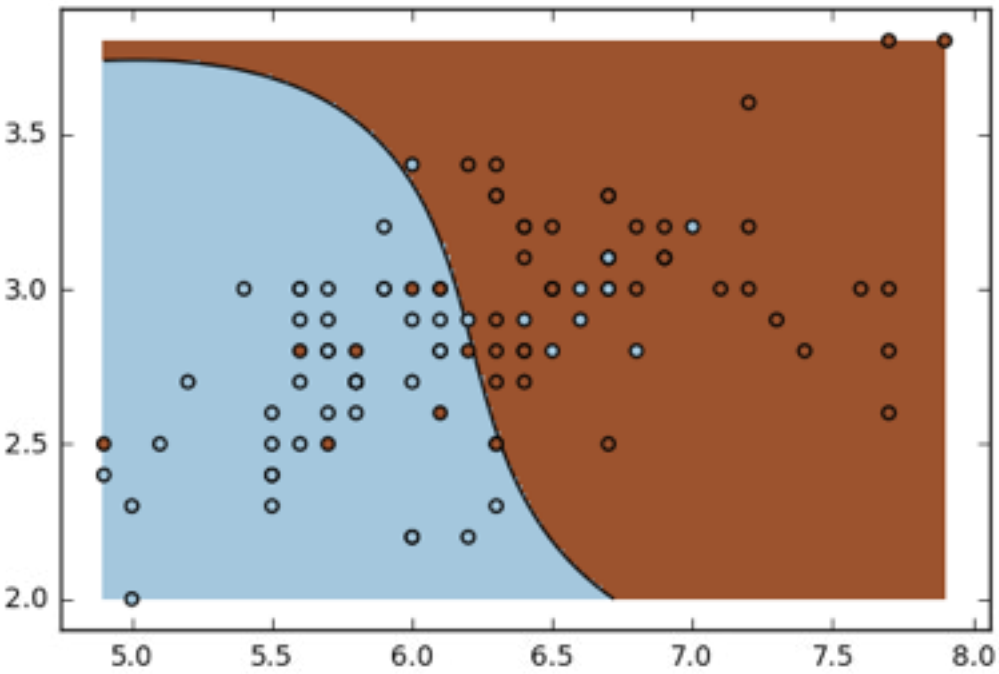
- A model that determines how the computer approaches the problem of assessing the data, with parameters that fine-tune the model to make the predictions as accurate as possible. There are numerous machine learning models.

- The cost function that calculates the error, or how far the algorithm is from perfect performance.

A description of one supervised machine learning model, the support vector machine (SVM), can illustrate how machine learning works on a more technical level when applied to the problem of predicting whether an email is spam or not spam. Starting with a dataset of emails (represented as dots on the graph on the next page) labeled as spam or not spam, an SVM will try to find the line that best separates the two categories of data points, as seen below with the line dividing the blue and brown areas.[3] The parameters in this case determine the location and curves of the line. The SVM distinguishes between the categories through an iterative process of

3 The idea of this chart was drawn from Scikit-learn. For more, see Fabian Pedregosa et al., 'Scikit-Learn: Machine Learning in Python', *Journal of Machine Learning Research* 12, no. Oct (2011).

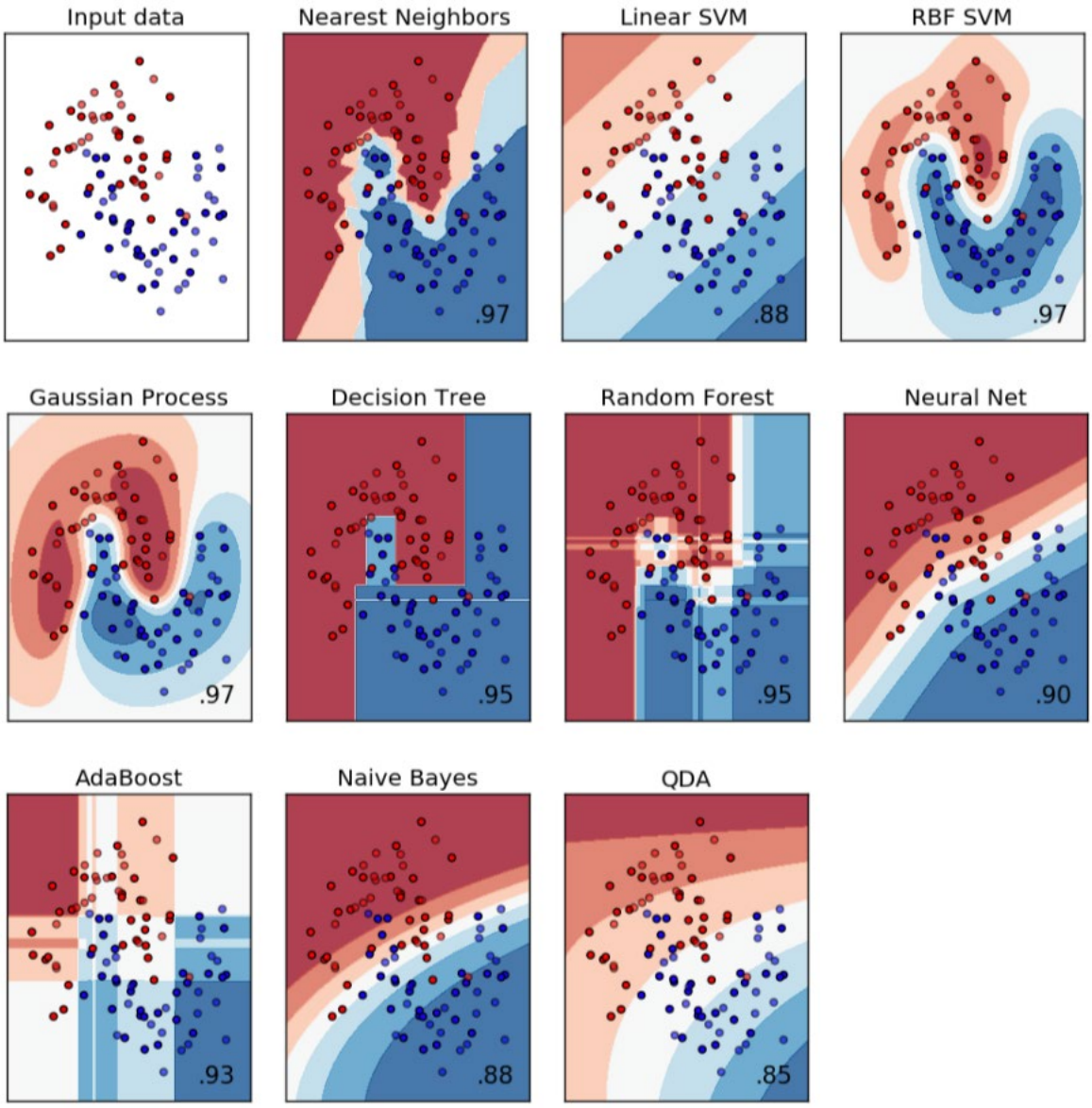
starting with a random line, determining what changes to the line (made through parameter adjustments) decrease its error as calculated by the cost function. It then iteratively adjusts its parameters until the error is minimized to its fullest extent; at that point, the best separating line has been found. Once that line is found, new emails can be plotted on this graph and their category (blue or brown; spam or not spam) can be predicted from their relationship to this line.[4]



SVM is only one of numerous supervised learning models that are used, and a number of models take different approaches to the same problem. The image on the next page shows how ten different supervised learning models divide the same set of data into two categories, red or blue; the shading indicates confidence intervals.[5] Importantly, each algorithm has its strengths and weaknesses, and machine learning practitioners will often try a number of models to determine which one works best for the problem they are trying to solve.

4 It is worth mentioning that while the illustration below only shows two dimensions, reflective of examining two characteristics of the data, the separating line in most machine learning cases is actually a multidimensional hyperplane.

5 This image is drawn from Scikit-learn. Pedregosa et al., 'Scikit-Learn: Machine Learning in Python.' For the specific original image, see http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py



A common analogy is that the supervised learning model is a box with thousands of adjustable knobs, which represent parameters, and the goal of supervised learning is to adjust the knobs to find the configuration that minimizes overall error. Adjusting the parameters allows the model to uncover the patterns in the data that are important for prediction. After the model finds the best parameter configuration, it can use this configuration to make predictions when given new data. In addition to support vector machines, there are a number of different types of supervised learning algorithms, such as decision trees, Bayesian networks, and more. The details of each of these are beyond the scope of this paper. More important is that, while these approaches are diverse in their applications and relative advantages, they all follow the same basic concept.

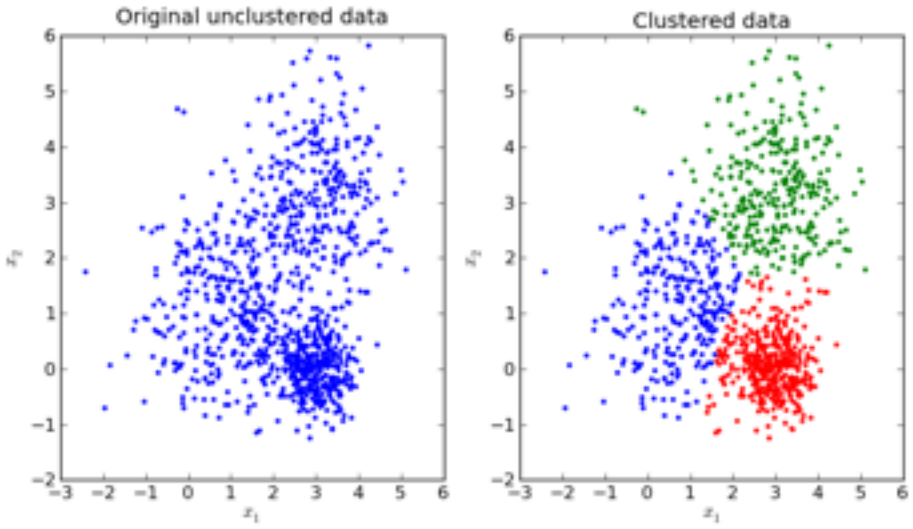
Unsupervised Learning

Supervised learning algorithms benefit from the guidance provided by the training set of data, and sometimes by real-time feedback about whether their predictions are correct. Sometimes data isn't so neatly structured, though. Unsupervised learning is more useful when there is not a clear outcome of interest about which to make a prediction or assessment. Unsupervised learning algorithms are given large amounts of data and try to identify key structures, or patterns, within them.

One common task for these algorithms is to spot clusters in a set of data. The clusters represent groups that each share meaningful characteristics. Clustering is very useful in market segmentation, for example. It can break data representing an undifferentiated sea of customers into groups that share preferences and interests, enabling companies to better tailor their products and marketing to each group. But a key insight bears repeating: in unsupervised learning, the algorithm finds the clusters on its own, and is not given any preconceived notions about how to break down the data into groups.

One unsupervised machine learning method, k-means, can help illustrate how unsupervised learning works. A company may want to perform market segmentation with the customer data shown in the image, dividing its customers into three segments for more accurate advertising or pricing. The k-means algorithm would randomly propose three (or the desired number of clusters) points on the graph to be the centers of the new clusters. It would then adjust these center points iteratively in the direction that minimizes the distance between the center point and all of the points in its cluster, while also maximizing the distance between the center point and all points not in its cluster. The end product, as shown in the image, is three separate and well defined clusters, with customers matched to the cluster of the closest center point.[6]

6 For more, and for the original image, see Nathan Landman et al., 'K-Means Clustering', Brilliant.



One of the most common uses of unsupervised learning is to better understand the structure of data in order to build better supervised learning algorithms. For example, unsupervised learning can be used to combine the multitude of pixels from a picture into a small number of important recognizable features. These features, such as the structures of the eyes, nose, and mouth can then serve as an input for a supervised learning facial recognition algorithm. Notice that the each of the three images below highlights facial features that would be important for identification—features derived through unsupervised learning.[7]

7 For more, see Pedregosa et al., 'Scikit-Learn: Machine Learning in Python.' For the original image, see http://scikit-learn.org/stable/auto_examples/decomposition/plot_faces_decomposition.html#sphx-glr-auto-examples-decomposition-plot-faces-decomposition-py

Reinforcement Learning

Rather than simply manipulating data, reinforcement learning algorithms work by introducing software known as a machine learning agent to an environment and teaching it how to act. Unsurprisingly, reinforcement learning is very important in robotics, though its most public successes have been in defeating humans in games.

DeepMind, a leading artificial intelligence company that was purchased by Google in 2014, has been a pioneer in reinforcement learning. In 2015, they devised a program that would learn how to play basic video games. The agent in this program got only the information on the screen, including the score of the game, and nothing else—not even the rules. It then proceeded to make decisions, randomly at first, and to see the rewards or failures of its choices. Over hundreds or sometimes thousands of iterations of the game, the agent saw which decisions, or series of decisions, led to better rewards in certain conditions. This information became the basis of its approach to playing the games well. The DeepMind game-playing agent was able to beat professional game players in more than three-fourths of the games it tried.[8]

Reinforcement learning returned to the news in 2016 with perhaps its highest profile success. DeepMind applied reinforcement learning alongside deep learning (discussed below) to the ancient board game Go. Go is a game that had long been considered too difficult for artificial intelligence to master because of the many combinations of possible moves; there are many more possible Go board combinations than atoms in the universe, and many more combinations than in even other complex games like chess. Playing Go well is not just a matter of calculation, but of intuition—something at which machines are famously weaker than humans.

DeepMind's agent, named AlphaGo, observed many thousands of games of professional Go to understand the important patterns. It then began to play millions of games of Go against itself, refining its capabilities and uncovering additional insights. This highlights the immense scale made possible

8 Dharshan Kumaran and Demis Hassabis, 'From Pixels to Actions: Human-Level Control through Deep Reinforcement Learning', Google Research, 25 February 2015.

by machines; AlphaGo in a short period of time played more games of Go than even the most dedicated players can play in a lifetime. It used the insights from these games in a widely-publicized set of matches in 2016 and 2017 to defeat the top players in the world. In so doing, it introduced new ideas and strategies that had eluded players ever since the invention of the game.[9]

9 Cade Metz, 'In Two Moves, AlphaGo and Lee Sedol Redefined the Future', Wired, 16 March 2016.

The Current State of Machine Learning

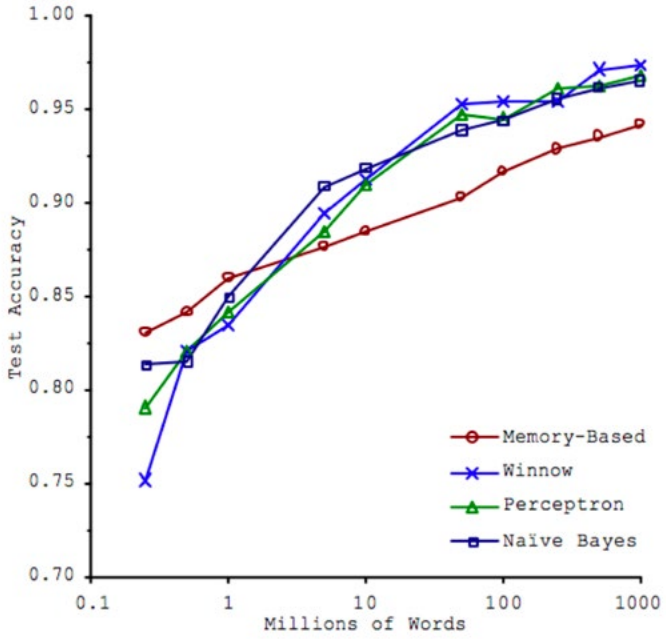
An old joke goes something like this: "Once something works, we stop calling it artificial intelligence and start calling it software." Moving the goalposts of artificial intelligence, of which machine learning is a part, is a repeated pattern: as advances have continued, such as defeating humans in chess, then *Jeopardy!*, and then later in Go, we have been reluctant to recognize each step as artificial intelligence in and of itself. Instead, artificial intelligence is often viewed as something that is perpetually over the horizon. While there are many goals that are out of immediate reach, this perspective can obscure the progress that has already been made. By focusing on the ways in which machine learning has already been applied with success and looks poised to grow still further, this section draws out several important and overlapping concepts of critical importance.

The Importance of Data

While conceptual breakthroughs in the design of machine learning programs are significant, machine learning still relies on data. In a groundbreaking 2001 paper, Michele Banko and Eric Brill showed that the amount of data used to train machine learning algorithms has a greater effect on prediction accuracy than the type of machine learning method used; see the graph on the next page.[10] In other words, for some problems, a decent algorithm that learns from a lot of relevant data outperforms a great algorithm that learns from minimal or poor data. This is one of the reasons why some of the most successful companies today are the ones that have the most data on which to train their programs, and why companies are willing to pay massive amounts of money for more data. As Peter Norvig, Google's Chief Scientist, once said, "We don't have better algorithms than anyone else; we just have more data."[11]

10 Michele Banko and Eric Brill, 'Scaling to Very Very Large Corpora for Natural Language Disambiguation' (paper presented at 'Proceedings of the 39th Annual Meeting on Association for Computational Linguistics', 2001).

11 Scott Cleland, 'Google's "Infringenovation" Secrets', Forbes, 3 October 2011.



But high-quality data collection can be notoriously difficult and expensive. Google and Facebook gained success in part by finding cheap ways to collect data from their users, which improves their services and encourages users to make still more data available. Unfortunately, many public policy problems do not have large amounts of accessible data from which to learn. Social welfare programs can be inefficient at providing the right resources to the right people at the right time, but there often is precious little data on the details of this inefficiency. Some data, for example, can only be collected by conducting surveys in person, which is cost-prohibitive and time consuming. Other data, such as medical records, are kept private by default.

The Deep Learning Approach

Deep learning is perhaps the most promising area of machine learning today.[12] While supervised, unsupervised, and reinforcement learning are all overarching methods, deep learning is an architecture that can implement those methods; for example, deep reinforcement learning systems are quite powerful. Deep learning uses networks that contain layers of nodes that in

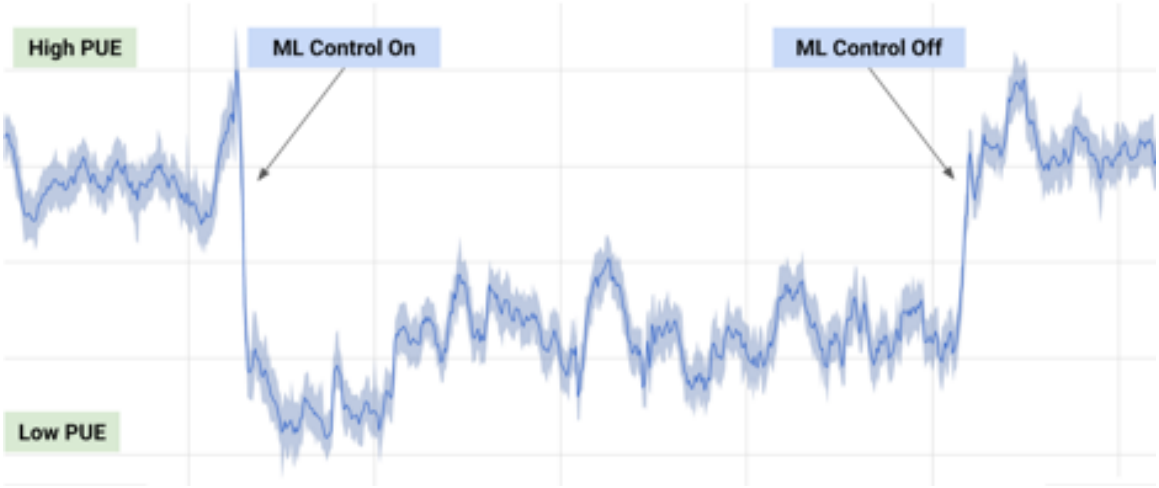
12 For a more detailed overview, see Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (MIT Press, 2016).

some ways mimic the neurons in the brain. Each layer of neurons takes the data from the layer below it, performs a calculation, and provides its output to the layer above it. Deep learning can combine an unsupervised process to learn the features of the underlying data (such as the edge of a face) and then provide that information to a supervised learning algorithm to recognize features as well as the final result (correctly identifying the person in the picture). More generally speaking, deep learning is useful for capturing hierarchical meaning; for example, it can grasp from images that cats have body parts, and not the other way around, and that those body parts are made up of shapes. Deep learning is used today for everything from better understanding the molecular interactions inside human cells to improving computer vision and natural language processing.

Another area in which deep learning has yielded tangible benefits is in improving energy efficiency. For example, the operation of thousands of servers that drive any tech company creates a great deal of heat. As a result, cooling data centers is an enormous challenge and one that, if not managed well, can quickly become financially and environmentally costly. While Google had already made enormous efforts to increase its cooling efficiency—from 2011 to 2016, the company tripled how much computing power it got per unit of energy—machine learning enabled still greater progress.

DeepMind devised an approach that improved cooling efficiency in Google's data centers by 40 percent. The company trained neural networks with historical data from thousands of sensors to understand the complex interactions between equipment, operational decisions, and environmental factors such as weather. In so doing, the machine learning approach was able to identify subtle but substantial ways in which even a company as advanced as Google could improve.[13] Google's graph provides a visualization of how enabling the machine learning approach reduces power usage (as measured by a Google statistic known as PUE; lower is better). In a world warming due to climate change and facing still growing energy demands, this example provides hope that machine learning could help meet energy needs through increased efficiency.

13 Richard Evans and Jim Gao, 'DeepMind AI Reduces Google Data Centre Cooling Bill by 40%', DeepMind, 20 July 2016.



Progress in Computer Vision

Machine learning enables computers to identify objects in pictures. This skill can be used to make predictions that exceed human accuracy. As outlined earlier, to achieve computer vision, the first layer of a deep learning program is given the data from individual pixels in an image, and it learns which characteristics are most important. At the lowest levels of the process, the characteristics or features identified are as simple as finding the edge of an object. The features are progressively passed up to higher layers in the network, where more complex features are learned. The final layer then uses these features to identify the objects in the picture.

This capacity for computer vision has a wide range of applications, most notably in medicine. For example, it is often difficult to develop accurate prognoses for lung cancer patients. Using several thousand images, however, a machine learning algorithm was able to examine many detailed and nuanced characteristics of the cancers. This went beyond size and shape of cell to include things like spatial relationships between cells and the texture and shape of cell nuclei. While human experts regularly examine several hundred characteristics of cancer cells in order to make a prognosis, machine algorithms were able to examine nearly ten thousand characteristics and determine which were most important. As one investigator noted, "the computers can assess even tiny differences across thousands of samples many times more accurately and rapidly than a human."[14] The result is better assessment of patients' conditions and better understanding of the

14 Yun Liu et al., 'Detecting Cancer Metastases on Gigapixel Pathology Images', *arXiv preprint arXiv:1703.02442* (2017).

progression of cancer; other examples of deep learning algorithms can aid pathologists and reduce human errors by 85 percent.[15]

Improvements in Natural Language Processing

One common aspiration, imagined in a plethora of science fiction works, is a computer that can read, listen, understand, translate, and talk. Formally, this kind of work is known as natural language processing. Machine learning has enabled significant advances in this area. Google Brain, a machine learning division of the company that focuses on applications to Google products, had enormous success in improving the quality of translations when it deployed a neural network-based approach. In one of its early deployments, the machine learning algorithm was able to improve Google Translate's French BLEU score—a key metric for evaluating translation quality—by seven points. Previously, improvements of one and two points were considered impressive.[16]

More work remains to be done on natural language processing. Andrew Ng, the founder of Google Brain and the former Chief Scientist at Baidu, estimates that computers can recognize about 95% of speech in 2017.[17] Ng believes that while the rate of mistakes is reasonably low, it is still significant enough to pose a substantial hurdle in interactions; he thinks the difference between 95% and 99% accuracy is the difference between talking to computers sporadically as we do today and seamlessly talking to computers without thinking anything of it.[18] There are many challenges in this area, such as getting computers to understand ideas rather than just transcribe or act on them. If those problems are solved, the nature of human interaction with machines will be very different.

15 Dayong Wang et al., 'Deep Learning for Identifying Metastatic Breast Cancer', *arXiv preprint arXiv:1606.05718* (2016).

16 Gideon Lewis-Kraus, 'The Great A.I. Awakening', New York Times, 14 December 2016.

17 Rebecca Merrett, 'Future of Mobile, IoT Driven by Speech Recognition: Andrew Ng', CIO, 6 May 2015.

18 Andrew Ng, 'AI: The New Electricity', YouTube, 11 June 2016.

The Ever-Increasing Internet of Things

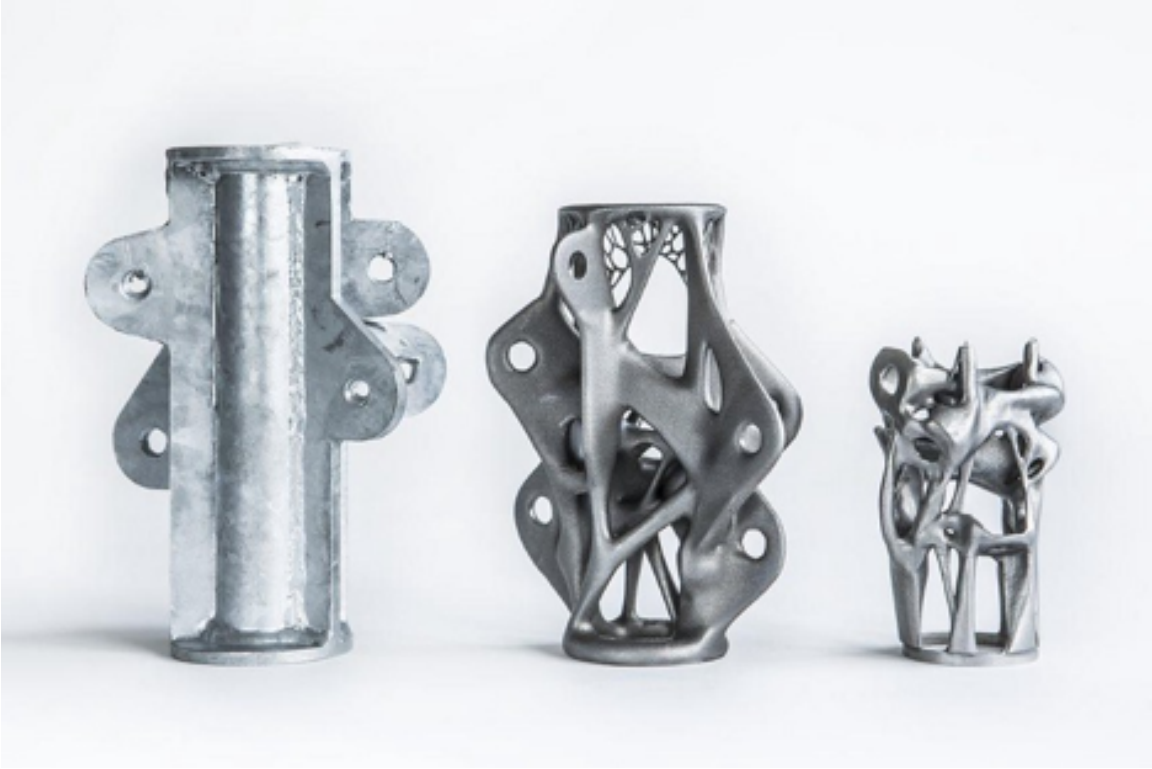
The growth in machine learning intersects with massive growth in other areas. Perhaps chief among these is the Internet of Things—interconnected devices of all types, from thermostats to toasters. Machine learning provides the potential to allow each device to learn its user's personal preferences and proactively get better at its task. For example, each Nest-branded thermostat gathers information about its user's habits and temperature preferences, and eventually learns to set the temperature to optimal levels itself, based on a variety of factors. This can result in gains in not just comfort, but also energy efficiency.

Toasters and thermostats may, in the scheme of life, be of reasonably low significance compared to near-future Internet of Things devices. Self-driving cars, for example, will use machine learning to stay on the road, but also to customize themselves to their various users or to best predict which route is fastest. A wide range of companies will employ machine learning in their equipment, sometimes referred to as the Industrial Internet of Things. Some of this equipment, such as the software and hardware that makes up the industrial control networks of critical infrastructure, serves vitally important societal functions. While there are no doubt enormous gains in efficiency to come from the convergence of the Internet of Things and machine learning, there is also no doubt about the stakes. In situations like these, machine learning algorithms must perform exceptionally well or risk dangerous consequences.

A Changing Approach to Design

What enables machine learning algorithms to achieve better results on some problems? It is not a question of machines being "smart" or humans being "dumb." Instead, improvements are often driven by machines' ability to execute a large number of calculations and better account for enormous amounts of data. More provocative, perhaps, is the idea that machines are not wedded to conventional notions. Instead, they will consider (and sometimes use) approaches that are counterintuitive but superior. Taken together, the ability to consider many possibilities and the capacity to be

free of preconceived notions is powerful.[19] It also results in some solutions that seem alien in nature, even though they work quite well. For example, the image below shows how a generative design algorithm—a computer-driven approach to design—devised an unconventional load-bearing column that uses vastly less material but is equally effective. This capacity for efficient design holds enormous promise in areas like manufacturing.[20]



19 For more, see Kristina Shea, Robert Aish, and Marina Gourtovaia, 'Towards Integrated Performance-Driven Generative Design Tools', *Automation in Construction* 14, no. 2 (2005).

20 '3D Makeover for Hyper-Efficient Metalwork', Arup, 11 May 2015.

The Transformational Effects of Machine Learning (and Their Challenges)

Andrew Ng, the aforementioned research scientist formerly of Google Brain and Baidu, provides a useful heuristic for understanding where machine learning is today: "If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future."[21] He has called AI "the new electricity," in that it at some point will be ubiquitous and will fundamentally reshape how humans live and societies function.[22] While the previous section showed how machine learning has already made substantial improvements in a range of areas, this section highlights the ways in which machine learning, now and especially in the future, will present new opportunities and challenges in particular areas vital for policymakers.

Before that, however, it is worth reflecting on the limits of machine learning. These, in some ways, can be counterintuitive: machine learning algorithms get better at some tasks much faster than they get better at others. This is often the case even if other types of non-machine learning software perform differently, even if humans learn both tasks at the same speed, and even if humans think that both tasks are equally part of "intelligence."[23] For example, machine learning algorithms are often much better at recognizing parts of images than they are recognizing how words relate to concepts. They are also, as a generalization, much worse at learning when they have less data on which to rely; large amounts of data often overwhelm humans, but often have a positive, rather than negative, effect on machine learning algorithms' performance. Similarly, while humans are very good at quickly transferring ideas from one context to another, machines sometimes struggle with this. As we consider how machine learning affects specific disciplines, we have tried to highlight the tasks in those disciplines in which algorithms are likely

21 Andrew Ng, 'What Artificial Intelligence Can and Can't Do Right Now', Harvard Business Review, 9 November 2016.

22 Ng, 'AI: The New Electricity'.

23 For more on the expected rate of improvement in AI across different tasks, see Katja Grace et al., 'When Will AI Exceed Human Performance? Evidence from AI Experts', ArXiv, 30 May 2017.

to be very effective and impactful, but it is worth remembering that algorithms will not be universally so.

Future War

Many research scientists and corporations building artificial intelligence have pledged not to work on projects related to weapons.[24] Nonetheless, machine learning is of enormous interest to military planners and strategists. In the United States, it is viewed as central to the so-called "Third Offset"—the way in which the American military will retain superiority over other nations (the first two offsets were nuclear weapons and precision guided munitions). Artificial intelligence, the thinking goes, will enable the United States to field better weapons, make better decisions in battle, and unleash better tactics. United States Deputy Secretary of Defense Bob Work sums it up, saying, "I am starting to believe very, very deeply that it is also going to change the *nature* of war."[25] For example, semi-autonomous swarms of aircraft might be more capable of carrying out certain objectives than individual human pilots or might be more survivable in contested airspace, and would thus give a decisive edge to a nation that had them over a nation that did not.[26]

As ever, the application of machine learning requires data, and sometimes data that is hard to get. William Roper, the head of the Pentagon's Strategic Capabilities Office, highlighted the role of gathering information for machine learning in times of conflict. He said, "It's wealth *and* fuel. Your data keeps working for you. You stockpile the most data that you can and train that to teach and train autonomous systems." Roper said that in future military engagements, "the purpose of the first day or the second day will not be to go out and destroy enemy aircraft or other systems. It's to go

24 'Autonomous Weapons: An Open Letter from AI & Robotics Researchers', Future of Life Institute, 28 July 2015.

25 Sydney Freedberg, 'War without Fear: DepSecDef Work on How AI Changes Conflict', Breaking Defense, 31 May 2017. Emphasis in the original.

26 Robert O. Work and Shawn Brimley, '20yy: Preparing for War in the Robotic Age', Center for a New American Security, January 2014. Kareem Ayoub and Kenneth Payne, 'Strategy in the Age of Artificial Intelligence', *Journal of Strategic Studies* 39, no. 5-6 (2016).

out, collect data, do data reconnaissance, so that our learning system gets smarter than [the enemy's]."[27]

With the right data, machine learning can change how a nation prepares for and fights wars. Generative design algorithms, for example, can affect how nations build and deploy military technologies before battle begins. Former DARPA Director Arati Prabhakar highlighted ways in which machine learning can change how militaries fight once conflict starts. Electronic warfare, which has traditionally been a slow-moving area of engagement in which each side carefully studies the technology of the other and then determines how to jam it, is ripe for such a change. "We want to get to where we respond and react faster than human timescales," she said. "The way we do that is by, first of all, scouring the [electromagnetic] spectrum in real time and, secondly, applying some of the most amazing frontiers of artificial intelligence and machine learning, techniques like reinforcement learning. [Then we] use those to build systems, onboard systems, that can learn what the adversary is doing in the electromagnetic spectrum, start making predictions about what they're going to do next, and then adapt the onboard jammer to be where the adversary's going before they get there."[28]

Machine learning in a warfighting context raises some potentially serious and unique issues. These are in addition to the general challenges related to machine learning discussed below. The role machines might have in deciding to carry out a lethal strike is of foremost concern. The Obama Administration urged caution when it came to developing and deploying autonomous lethal weapons.[29] However, it is not difficult to imagine how requiring human involvement in some parts of conflict could slow down decision-making in a way that is unacceptable to some military planners: for example, when it comes to countering cyber attacks that occur very rapidly, or if an adversary is employing machine learning-based warfighting technologies of their own. Unsurprisingly, a great number of military

27 Patrick Tucker, 'The Next Big War Will Turn on AI, Says US Secret-Weapons Czar', Defense One, 28 March 2017.

28 Sydney Freedberg, 'Faster Than Thought: DARPA, Artificial Intelligence, & the Third Offset Strategy', Breaking Defense, 11 February 2016.

29 'Preparing for the Future of Artificial Intelligence', Executive Office of the President: The White House, 2016.

and cybersecurity strategists advocate for more automation.[30] Perhaps machines will also some day be both faster and better than humans at certain decisions—in a lot of areas, such as some forms of anti-aircraft defense, this threshold has already been crossed. While we appreciate the value such automation offers to governments, we stress the vital importance of considering the ethics and impacts first. This sort of analysis is impossible without a firm understanding of what exactly machine learning entails and what its limitations are.

Healthcare

The healthcare industry makes up almost a fifth of the American economy and is home to a great deal of machine learning-driven innovation. This is a trend that will grow, with improvements coming in prediction, computer vision (as discussed above), personalized medicine, and drug discovery.

Broadly speaking, the American healthcare system is in the midst of a transformation from fee-for-service medicine, where providers are paid for each additional treatment, to value-based care, where providers are paid for keeping patients healthy.[31] The result is an environment where doctors and hospitals are now incentivized to advance overall patient health; it is cheaper to keep patients healthy proactively than to admit them for expensive hospital care later. This renewed emphasis on prevention necessarily relies on prediction, because doctors must identify which patient is likely to develop which serious medical problem and determine how best to prevent it. Medical providers are starting to use prediction algorithms to hypothesize who is likely to come back to the hospital with a complication after being discharged, which hospital patients are at risk for developing

30 See above for a sampling of views on speed and automation in military conflict. For more on speed and automation in cyber operations, see Richard Clarke and Robert Knake, *Cyberwar* (New York: HarperCollins, 2010), 30. Joel Brenner, *Glass Houses* (New York: Penguin, 2014), 199. 'Budget Request for Information Technology and Cyber Operations Programs: Written Testimony of Gen. Keith Alexander', Committee on Armed Services: US House of Representatives, 2012. Ben Buchanan, *The Cybersecurity Dilemma* (New York: Oxford University Press, 2017); Ben Buchanan, 'The Legend of Sophistication in Cyber Operations', Belfer Center for Science and International Affairs, January 2017.

31 For more, see Michael E Porter and Elizabeth Olmsted Teisberg, *Redefining Health Care: Creating Value-Based Competition on Results* (Harvard Business Press, 2006).

sepsis, and how to best allocate resources and health care providers in the community to keep patients healthy.

Prediction of individual patient outcomes also enables more personalized medicine, one of the most exciting areas of machine learning-led innovation. The more individualized treatment can be, the more it can recognize and respond to the varying needs of specific patients. For example, cancer can develop from numerous different mutations, and the effectiveness of chemotherapy depends heavily on the exact fingerprint of mutations in each tumor. Machine learning enables researchers to discover important mutations and better target drugs.[32]

Machine learning also has the potential to find new cures for diseases. Some companies, such as Sema4 and Flatiron Health, are using machine learning to look through massive amounts of medical and genomic data to better target existing treatments and look for new targets for drugs. Sema4 is looking through thousands of individuals' genetic information to find people who are genetically predisposed to rare genetic diseases but are not affected. The company hopes to find what other mutations may be protecting these people. The protective mutations may be the key to finding drugs for curing many rare diseases.[33]

A final point deserves mention: some of machine learning's most significant impacts on Americans' health may come not from medical changes but transportation ones. More than 30,000 Americans die in car accidents each year; to the extent that self-driving cars can reduce the number of crashes, they can directly save American lives.[34] On the other hand, a reduced number of car-related deaths also reduces the number of organs available for transplants—a negative side-effect to a development that is much more positive overall and that can be countered by policies to encourage organ donation or by advances in 3D printing of organs. It suffices to say that the intersection of the healthcare system

32 For two overviews of this topic, see Margaret A. Hamburg and Francis S. Collins, 'The Path to Personalized Medicine', *New England Journal of Medicine* 363, no. 4 (2010). Michelle Whirl-Carrillo et al., 'Pharmacogenomics Knowledge for Personalized Medicine', *Clinical Pharmacology and Therapeutics* 92, no. 4 (2012).

33 Mark Warren, 'The Cure for Cancer Is Data—Mountains of Data', Wired, 19 October 2016.

34 'Fatality Analysis Reporting System', National Highway Traffic Safety Administration: Department of Transportation, 2017.

and machine learning is complex and multi-faceted. There is great room for improvement and innovation, but this requires careful foresight and well-structured systems for managing and sharing data, as well as for considering the broader social impact.

Law Enforcement

Many public service agencies will likely find machine learning to be of substantial value in more efficiently allocating limited resources. Some of these organizations, such as law enforcement entities, have not taken advantage of advances in machine learning due to concerns about cost and complexity. The agencies that have already adopted machine learning, however, have found that it helps them make better use of their available resources.

The Santa Cruz Police Department (SCPD) in Santa Cruz, California has adopted machine learning techniques for predictive policing. The SCPD uses a real-time predictive algorithm to identify the fifteen 150x150 meter squares in the city of Santa Cruz that are at highest risk for crime at any one time. The department shares this information with its officers, who can then devote more of their shift time to these areas. One important aspect to this policy is that the algorithm is not attempting to replace the judgment of police officers on the street or give orders. Instead, the algorithm generates a simple prediction that is one of many factors that officers use when deciding where to patrol. One crime analyst with the SCPD described predictive policing as similar to an algorithm that tells fishermen where fish are; it can help them be better at their jobs, but it can't replace them. Santa Cruz saw a fourteen percent drop in burglaries after the algorithm was implemented. The department finds predictive policing especially useful because they have experienced significant staff cutbacks and are expected to do more with less—a trend familiar to police organizations across the country.[35] Machine learning will be used more and more as the engine of these sorts of data analytics efforts.

35 Emma Pierson, Sam Corbett-Davies, and Sharad Goel, 'Fast Threshold Tests for Detecting Discrimination', *arXiv preprint arXiv:1702.08536* (2017).

Data-driven policing is not new, and predictive policing has become controversial due to its potential to target minority neighborhoods. The NYPD CompStat program, begun in the mid-1990s, is a massive initiative to use data-driven policing to lower crime. The effort initially raised concerns about bias in the program that led to greater effects on minority groups. Additionally, the department's reliance on data also incentivized some officers to manipulate numbers and focus on measured results, rather than the needs of their communities.[36] The NYPD has made significant changes to the program, and introduced CompStat 2.0, which, in order to increase transparency, gives the public access to all of the information the police use.

Going forward, machine learning can be seen as a force multiplier for police departments, allowing them to be more effective with limited resources. But data-driven approaches are not panaceas. The CompStat case shows that a more rigorous analysis process can be narrowly more effective, but may not necessarily accomplish the organization's wider goals if it is used to implement a problematic underlying policy; the matter of bias and machine learning, which extends beyond law enforcement, will be discussed in more detail in the next section.

Body cameras are another area of significant interest at the intersection of machine learning and policing. As more and more police wear body cameras, more data is available to be collected. Machine learning algorithms could use this data for a variety of purposes with a range of social effects. The particulars of how this data is controlled, collected, and analyzed deserve attention.

36 Chris Francescani, 'NYPD Report Confirms Manipulation of Crime Stats', Reuters, 9 March 2012.

General Policy Challenges Posed by Use of Machine Learning

Machine learning has been around for a number of decades, but as its use continues to accelerate, policymakers must confront the broad challenges it poses. Machine learning is now demonstrably relevant to the realm not just of technology or science, but also of wider public policy. Algorithms already exist that can read our medical records, decide if we qualify for a loan, identify our political beliefs, and recommend a prison sentence. The machine learning technology community has already started taking many of the concerns outlined in this section seriously; it is time for the policy community to think deeply and in an informed way about these challenges as well. Two significant reports by the Obama administration in 2016 represent a good start in this direction, but much more remains to be done.[37]

Data Availability

That we live in a world of big data may be a cliché, but it is also a fact. Most of the data generated throughout human history has been been generated in recent years. This data, combined with machine learning, has a massive potential to change our lives, from poverty elimination to healthcare to environmental protection. Yet, this data is fragmented and mostly unreachable. Private companies or other organizations hold much of it, and even publicly available data can be extremely difficult to access and combine in any useful way. The problem of data availability is one of the areas where policymakers can help drive innovation in machine learning by encouraging open data initiatives.

Healthcare data provides a useful case study of the challenge of promoting data availability. The American healthcare system generates massive amounts of data, and access to this data could drive life-saving research. As mentioned above, some companies and researchers are currently trying to use that data to find better treatments for disease and improve patient

37 'Preparing for the Future of Artificial Intelligence', Executive Office of the President, 2016. 'The National Artificial Intelligence Research and Development Strategic Plan', National Science and Technology Council: The White House, 2016.

outcomes. However, this data is notoriously fractured, with different hospitals using different electronic health record (EHR) companies and little ability to share or combine data. Even hospitals that combine into unified healthcare systems cannot get their EHRs to share data seamlessly. The result is that if a patient has laboratory tests or imaging done at one hospital, then visits another hospital, the second hospital often cannot see the results from the first. On top of affecting patient care, this makes it extraordinarily difficult to monitor quality or conduct research using data from a population larger than that served by one hospital, with the exception of several larger databases that are highly fragmented in their own right.

In 2008, fewer than one in ten hospitals even used basic EHRs. The HITECH Act, passed in 2009, attempted to solve the data availability problem by promoting the adoption and "meaningful use" of EHRs, meaning they could be used to operate clinically (such as in the prescription of medicines), monitor quality, and share data. Hospital use of EHRs increased from 9% in 2008 to 84% in 2015.[38] This itself is a massive success—paper records are much less common than ever before and data is more easily shared.

On the other hand, the idea of meaningful use of EHRs was highly controversial in the medical community. EHRs are expensive to install, require retraining, and change workflow. Providers complained that the reporting requirements were too demanding and expensive, and the requirements soured some physicians on the effort. Despite the widespread adoption of EHRs, the overall goal of integrated meaningful use has not been achieved; healthcare data is still highly fragmented by hospital and EHR company. In 2015, meaningful use was replaced with the Medicare Access and CHIP Reauthorization Act, which provided more flexibility to meet regulatory requirements. Andy Slavitt, the then-Administrator of the Centers for Medicare and Medicaid Services, acknowledged: "The meaningful use program as it has existed will now be effectively over and replaced with

38 Office of the National Coordinator for Health Information Technology. 'Non-federal Acute Care Hospital Electronic Health Record Adoption,' Health IT Quick-Stat #47. dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php. May 2016.

something better. We have to get the hearts and minds of physicians back. I think we've lost them."[39]

There is not the space here to parse the nuances of EHRs, but the example demonstrates the impact policymakers can have on data availability and the vital importance of finding the right policy option. Though the regulatory approach taken by the HITECH Act was unpopular, it was also highly effective at promoting the adoption of EHRs, an important though not sufficient step. Going forward, policymakers should focus on providing incentives and infrastructure for efficient information exchange, in healthcare and beyond.

Privacy

The amount of personal data needed for many machine learning applications seems to pose an inherent challenge to personal privacy. These algorithms can provide great value when they use individualized data to offer a degree of personalization. For example, Netflix can offer much better movie suggestions than cable television commercials can, but this is only because Netflix has access to a large amount of personal viewing history and the means to individually target its suggestions.

Suggesting binge-worthy shows may seem trivial, yet the same dynamic of individualized data yielding better results when given more intrusive access holds for more important predictions as well. As noted above, machine learning algorithms will soon be able to suggest the best treatment for a patient's specific tumor mutation fingerprint, but to do this they will need access to that patient's medical records. Privacy will continue to be a major challenge for machine learning policy, and protecting privacy is an area ripe for technical innovation in its own right.

A large and recurring threat to privacy stems from how easy it is to identify individuals in a supposedly de-identified dataset, or to use published algorithms to back-calculate information about the individuals the algorithm

39 Rajiv Leventhal, 'CMS's Andy Slavitt Says Meaningful Use Will Be Over in 2016', Healthcare Informatics, 12 January 2016.

trained on. This is especially easy when the information obtained can be combined with open source data. For example, Rui Wang and others used published Genome Wide Association Studies papers—which were supposedly reliant on datasets with personal details removed—to identify a number of the individuals used to train the algorithms. These researchers were able to uncover details about the individuals' genetic information.[40]

Likewise, Arvind Narayanan and Vitaly Shmatikov combined anonymized data that Netflix released as part of a machine learning competition with public information available on the Internet Movie Database website to identify users of the streaming site. The researchers noted that malicious attackers could potentially identify users' political beliefs and sexuality based on their movie preferences.[41] Similarly, after Massachusetts Governor Bill Weld assured the public that information made publicly available by the Massachusetts Group Insurance Commission was de-identified, Latanya Sweeney combined that data with voter registration data to identify the governor's medical records and mail them to his office.[42] In later testimony to the Department of Homeland Security (DHS), Sweeney testified that 87% of Americans can be identified using only their date of birth, gender, and ZIP code.[43]

The machine learning community has made promising advances that protect privacy but provide algorithms with the data they need. One area of significant progress is differential privacy, the concept that the output of an algorithm should not predictably change based on the presence or absence of any individual.[44] This ensures that the output of an algorithm cannot be used to learn information about any particular individual in the dataset. The technical details of these advances are beyond the scope of this paper, but promising general concepts include introducing statistical noise before

40 Rui Wang et al., 'Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study' (paper presented at 'Proceedings of the 16th ACM conference on Computer and communications security', 2009).

41 Arvind Narayanan and Vitaly Shmatikov, 'Robust De-Anonymization of Large Sparse Datasets' (paper presented at 'IEEE Symposium on Security and Privacy', 2008).

42 Nate Anderson, '"Anonymized" Data Really Isn't—and Here's Why Not', ArsTechnica, 8 September 2009.

43 'Statement of Latanya Sweeney', Privacy and Integrity Advisory Committee: Department of Homeland Security, 2005.

44 For one overview, see Cynthia Dwork, 'Differential Privacy: A Survey of Results' (paper presented at 'International Conference on Theory and Applications of Models of Computation', 2008).

training the algorithm, choosing an unknown subset of individuals and changing their characteristics before training an algorithm, and masking a random subset of data in a way that both ensures the algorithm does not conform to statistical noise and maintains differential privacy.[45] The overall concept of differential privacy has been effectively adopted by organizations ranging from the U.S. Census Bureau to Google and Apple as a means of gaining insight on populations without unduly affecting any individual's privacy.[46]

Other advances have helped protect privacy as well. In her testimony to DHS, Sweeney argued that post-9/11 America does not intrinsically face a choice between privacy and security. She believes machine learning can help to achieve both. She cited facial recognition technology she developed that uses unsupervised learning to combine features and ensure that any output will match a certain number of ambiguous individuals to protect privacy. She also developed a bio-terrorism surveillance algorithm that uses "selective revelation" protocols to only reveal identifiable medical information when absolutely necessary.[47]

It is worth noting, however, that these advances pertain only to algorithms trained on personal data, and not to the security of the data itself. Strong social norms maintain that the security and proper use of data is the responsibility of those who collect and access it, whether they are companies, academics, or governments. Perception of misuse of data can bring widespread condemnation, as Uber realized after announcing that it was tracking its customers' late night sexual encounters.[48] Data protection and use is a field in and of itself—one worthy of a great deal of attention—but advances in machine learning show that organizations that do secure and use their data responsibly can train valuable machine learning algorithms while protecting privacy.

45 This is usually called the holdout dataset. Cynthia Dwork et al., 'The Reusable Holdout: Preserving Validity in Adaptive Data Analysis', *Science* 349, no. 6248 (2015).

46 Ashwin Machanavajjhala et al., 'Privacy: Theory Meets Practice on the Map' (paper presented at 'IEEE 24th International Conference on Data Engineering', 2008). 'Apple Previews iOS 10, the Biggest iOS Release Ever', Apple, 13 June 2016.

47 'Statement of Latanya Sweeney', Privacy and Integrity Advisory Committee, 2005.

48 Douglas Perry, 'Sex and Uber's 'Rides of Glory': The Company Tracks Your One-Night Stands -- and Much More', The Oregonian, 20 November 2014.

Bias, Fairness, and the Misapplications of Machine Learning

On one hand, machine learning has the exciting potential to diminish many of the effects of bias on the day-to-day lives of Americans. Neutral algorithms could make hiring decisions, approve people for loans, and recommend criminal sentences without the implicit preconceptions that humans bring to the table. Properly implemented, this could lead to a more just society.

On the other hand, machine learning is sometimes critiqued as "money laundering for bias."[49] At its very worst, machine learning can cloak inequity with the imprimatur of science. A fundamental problem of applying machine learning to the real world is that a sophisticated algorithm is usually a "black box": while the user knows the prediction that is made, the process through which it is made is often far too complex to understand. Users often must simply determine if the algorithm's predictions are accurate enough to rely on given this constraint. If a trusted algorithm is in fact biased, it can further entrench unfairness and injustice.

This is not a theoretical risk, but one that has already manifested itself many times. For example, in a seminal case, a British medical school used an algorithm that screened out qualified female and minority applicants because it was trained on the decisions made previously by a biased admissions board.[50] A ProPublica investigation found that an algorithm developed by the company Northpointe, Inc. to provide a risk assessment score for judges during sentencing was racially biased and inaccurate. African Americans were much more likely to be labeled as high risk, but, as the table shows, the African Americans who were labeled as high risk were much less likely to commit another crime than high risk whites.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

49 Maciej Cegłowski, 'The Moral Economy of Tech', Society for the Advancement of Socio-Economics, 26 June 2016.

50 Stella Lowry and Gordon MacPherson, 'A Blot on the Profession', *British Medical Journal,* 296, no. 6623 (1988).

Much more needs to be done to study the problem. Almost no research has been conducted on how much these risk assessment scores influence sentencing decisions.[51] Furthermore, Northpointe has prevented access to its proprietary algorithm, further exacerbating the black box problem. This has caused significant concern, and justifiably so. Former U.S. Attorney General Eric Holder worried that the risk scores "may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society,"[52] while Supreme Court Chief Justice John Roberts said, "The impact of technology has been across the board and we haven't yet really absorbed how it's going to change the way we do business."[53]

There are a number of structural ways an algorithm can learn to be biased. First, it can train on data that doesn't represent the population to which it will be applied. Second, the algorithm can be trained to very accurately predict the results of an already biased system, as occurred in the flawed medical school admissions algorithm. When this occurs, the algorithm may not be introducing new bias into the system, but instead entrenching and masking it with a veneer of objectivity. Third, an algorithm can simply be trained poorly and learn prediction rules that are inaccurate or biased. Finally, a good algorithm can simply be applied in a way that introduces bias, as occurs when algorithms are inadvertently trained to overlook particularly small groups.[54]

All of this highlights an important concern: success in machine learning has traditionally revolved around prediction accuracy, and the most accurate models are considered the best. But it is important to understand that there is often a tradeoff between fairness and accuracy, as an algorithm may come to accurate conclusions that are unfair to certain groups.[55] For example, even if a particular algorithm is made more accurate by taking

51 Rose Eveleth, 'Does Crime-Predicting Software Bias Judges? Unfortunately, There's No Data', Motherboard, 18 July 2016.

52 Julia Angwin et al., 'Machine Bias', ProPublica, 23 May 2016.

53 Adam Liptak, 'Sent to Prison by a Software Program's Secret Algorithms', New York Times, 1 May 2017.

54 Claire Cain Miller, 'Algorithms and Bias: Q. And A. With Cynthia Dwork', New York Times, 10 August 2015.

55 Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes, 'A Confidence-Based Approach for Balancing Fairness and Accuracy' (paper presented at 'Proceedings of the 2016 SIAM International Conference on Data Mining', 2016).

racial factors into account, it seems unfair to use this information as a factor in making predictions because of the way it will prejudice judgment against certain groups. Indeed, the notion of what accuracy itself means for a given problem needs nuance; the algorithm that best approximates the decisions of a biased system is probably not the best algorithm for a just society.

As it has done with privacy concerns, the machine learning community has developed a number of approaches to tackle these problems. One approach would be for algorithms to identify the important features they use to make predictions. If someone is denied a loan, for example, the algorithm could provide an overall explanation. In this case, an acceptable answer would obviously have to be something like "not enough work history," instead of "not white." This sort of prose-based explanation is exceedingly diffcult for many complex machine learning systems to generate right now. One project that aims to address this, Locally Interpretable Model-Agnostic Explanations (LIME), attempts to use machine learning to break open the black box; DARPA has a project with similarly ambitious goals called Explainable AI.[56] These projects, if successful, could allow people to trust algorithms more as well as reveal any bias. A doctor, for example, would be more confident in an algorithm that could explain why it came to a certain diagnosis than in an algorithm that just suggested a diagnosis.

(a) Husky classified as wolf (b) Explanation

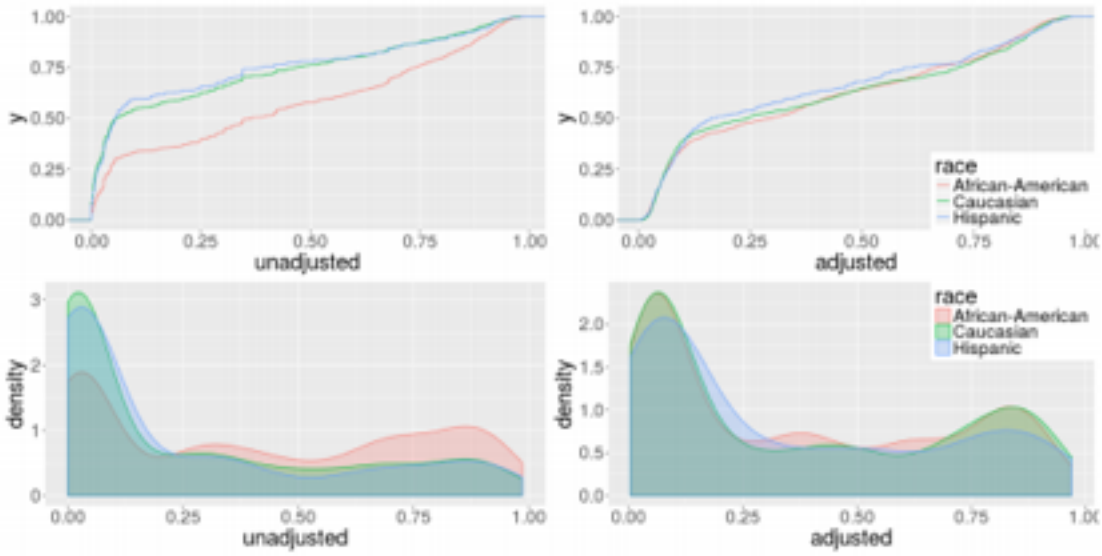
In making the case for their project, LIME's creators point to the example of an algorithm that was trained to distinguish huskies from wolves. Instead of relying on characteristics of the pictured animal, the algorithm instead learned to classify any picture with snow in the background as a wolf. The LIME explanation is necessary to understand that it was the background, and not the animal, that the classifier focused on, as the image shows.[57] While a snow-centric approach may have been effective for some

56 David Gunning, 'Explainable Artificial Intelligence', Defense Advanced Research Projects Agency.

57 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Why Should I Trust You?: Explaining the Predictions of Any Classifier' (paper presented at 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', 2016).

datasets, this demonstration shows the possibility of machine learning programs doing something fundamentally different in practice than their developers understand.

A number of machine learning researchers have attempted to solve the racial bias in sentencing problem as well. One paper proposes a more effective machine learning algorithm that can reduce crime by 25% without changing the jailing rate, or decrease the jailing rate by 42% with no change in crime, all while decreasing the percentage of African-Americans and Hispanics who are imprisoned.[58] Other researchers used machine learning methods to remove much of the racial bias from training data, and applied it to the Northpointe algorithm described above to decrease racial bias, as shown in the graphs.[59] The x axis for both graphs is the risk score, while the y axis in the top graphs is the actual reoffending rate and the y axis in the lower graphs is the amount of population with that score.



However, Cynthia Dwork, the computer scientist who introduced the aforementioned idea of differential privacy, argues that rather than attempting to hide characteristics like race from algorithms, data scientists should explicitly include them under a framework that ensures statistical parity. The exact meaning of statistical parity can be interpreted and calculated in a number of ways, but the general idea is to ensure fairness across

58 Jon Kleinberg, 'Human Decisions and Machine Predictions', (2016).

59 James E. Johndrow and Kristian Lum, 'An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction', *arXiv preprint arXiv:1703.04957* (2017).

categories, such as race. In such an arrangement, the demographics of the set of individuals who are given any particular classification are the same as the demographics of the broader population.[60] Interestingly, the same recommendation was made after the bias in the medical school admissions algorithm was discovered as well.[61]

Ultimately, we likely face a future where machine learning will make the problem of bias simultaneously better and worse. This challenge calls for greater input from the policy community. In the absence of guidance, some companies have set up internal ethics watchdogs in an attempt to police themselves. It is time for policymakers to ensure that machine learning increases fairness and does not entrench bias.

Economic Impact

Self-driving cars have attained a prominent spot in the American consciousness, even though fully autonomous vehicles are not allowed on the road. It has become almost a cliché to debate how the Trolley Problem (is it okay to kill one person in order to save five?) applies; should your self-driving car kill you in order to save a car full of passengers? While that is an interesting and perhaps important ethical debate, there are other more immediate ways in which machine learning can impact human society.

Consider a hypothetical boy named Sam who is born in 1975. Sam's father is a truck driver, and Sam from an early age develops a passion for cars. Sam is from the American heartland and, through public schools, attains a good education, though not a spectacular one. When Sam graduates high school in 1993, the internet is still used almost exclusively for obscure academic and defense pursuits; the first major browser company, Netscape, has not yet been founded. After a two-year associate's degree, Sam gets a job at his father's company as a truck driver, and very quickly demonstrates his skill. He is patient, conscientious, and an excellent mechanic for the many minor issues that crop up over thousands of miles on the road.

60 Cynthia Dwork et al., 'Fairness through Awareness' (paper presented at 'Proceedings of the 3rd Innovations in Theoretical Computer Science Conference', 2012).

61 Lowry and MacPherson, 'A Blot on the Profession.'

In 2017, Sam is 42 years old. He has been driving a truck for 22 years, and still loves it, even though the hours are long. He has two kids of his own, ages six and eight, and they're already better than he is with computers. Sam has started to hear about self-driving cars and trucks, and worries that they will replace him before too long. With projections of fully autonomous vehicles five to ten years away, Sam wonders if he should try to change careers, but he is afraid to give up his income at a time when his kids are young and his family needs the money. Yet he is concerned that by not switching he opens himself up to losing his job in a few years. He worries that one of the things harder than making a career change at 42 is making a career change at 52. He wishes he had realized in 1995 that truck driving was not a profession that would last, but how, in a mostly pre-internet age, could he have known?

What do we tell Sam in 2017? What does society owe him? Should the person who decided to go into computer software in 1995—which probably seemed like much less of a sure thing than truck driving—enjoy a long and incredibly lucrative career, while Sam worries about unemployment in middle age? Should Sam have acted to avoid the coming wave of automation, if not in 1995, then perhaps a decade or two later? Should we slow automation in order to protect Sam?

These are all questions to which we do not know the answer. Yet we believe they are questions of profound social importance. Most who have studied the issue agree: there will be more Sams, and not just in fields like truck driving. Nearly every occupation will be affected by machine learning. Terms like "the second machine age" or "the fourth industrial revolution" have been coined to describe the upheaval and change that will result.[62] These effects are not just limited to jobs that do not require a great deal of education. Indeed, some blue-collar careers, like hairdressing, are probably less likely to be affected than white-collar ones such as accounting.

Not all of these changes will necessarily result in unemployment.[63] Some jobs may involve tasks that are both easily automated as well as tasks that

62 For example, see Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (WW Norton & Company, 2014).

63 For more, see Clara Hendrickson and William A. Galston, 'Automation Presents a Political Challenge, but Also an Opportunity', Brookings Institution, 18 May 2017.

are very hard to automate. For example, a security guard might need to watch a dozen camera feeds looking for intruders (very easy to automate) but also to intercept the intruder (very hard to automate). This again is not merely a blue-collar phenomenon; a doctor might find that computers can replace her ability to read scans and monitor patients, but not her ability to interact with the patient or treat the disease.

Nor is the disruption of entire sectors bad in and of itself. Certainly, the internet of which Sam in 1995 was unaware has damaged a lot of industries, from travel agents to fax services, in such a way that has overall promoted an enormous amount of social good. A group known as the Luddites destroyed textile machinery because they thought it threatened employment; we do not suggest that machine learning be attacked for the same reason. But the data does indicate that there will be significant economic and social effects, and that those will have to be managed.

A series of papers shows the potential impacts, good and bad. In a 2016 report, two leading economists examined the concept of automation and concluded that it would offer substantial benefits to American workers. In particular, automation would supplant jobs in manufacturing, but it would offer opportunities to replace them with better ones in safer or more lucrative professions.[64] The same economists in 2017 examined data to better test their theory. They found that, thus far, it hasn't held up: the loss of manufacturing jobs didn't come with an offsetting growth in other areas.[65] One of the economists, Daron Acemoglu, drove home the stakes, saying, "The conclusion is that even if overall employment and wages recover, there will be losers in the process, and it's going to take a very long time for these communities to recover."[66] Sam, it seems, may be as archetypical as he is hypothetical—the implications of that remain to be seen.

64 Daron Acemoglu and Pascual Restrepo, 'The Race between Machine and Man: Implications of Technology for Growth, Factor Shares and Employment', 2016.

65 Daron Acemoglu and Pascual Restrepo, 'Robots and Jobs: Evidence from US Labor Markets', (2017).

66 Claire Cain Miller, 'Evidence That Robots Are Winning the Race for American Jobs', New York Times, 28 March 2017.

Security

Machine learning algorithms run on computers, and thus quickly intersect with the important matter of cybersecurity. There is an immediate concern: if systems employing machine learning are more powerful and more central to society, the potential for harm via hacking is much greater.[67] If machine learning algorithms are driving cars, fighting wars, and the like, not only are the stakes high, but the speed of individual decisions is likely to be high as well. Hackers who are able to compromise these systems thus have greater capacity to do enormous damage more quickly. Defenders might find it harder to intervene in time, and might find that the risk of cascading failures is harder to predict, with more damaging consequences. In short, traditional computer defense is of even greater priority for many machine learning systems.

In addition to these usual—and significant—concerns about data protection and securing systems from malicious intruders, machine learning presents some new challenges.[68] This is particularly relevant for algorithms that attempt to defeat an adversary that evolves, such as a spammer. A clever adversary may attempt to discover and exploit weaknesses in the machine learning program. This kind of operation is known as an exploratory attack. For example, if a spam-detecting machine learning program examined only the words in a message that also appeared in a dictionary, spammers may learn to slightly misspell words in order to evade detection. In the case of identification algorithms that rely on biometrics, imposters may seek to exploit the machine learning program's error tolerance in order to impersonate someone. These efforts undermine the integrity of the

67 For more on this principle and on how the harm can spread, see Thomas Rid and Peter McBurney, 'Cyber-Weapons', *RUSI Journal* 157, no. 1 (2012). Ben Buchanan, 'The Life Cycles of Cyber Threats', *Survival* 58, no. 1 (2016).

68 For more details and examples of some of the below challenges, see Ling Huang et al., 'Adversarial Machine Learning' (paper presented at 'Proceedings of the 4th ACM workshop on Security and artificial intelligence', 2011). Marco Barreno et al., 'Can Machine Learning Be Secure?' (paper presented at 'Proceedings of the 2006 ACM Symposium on Information, computer and communications security', 2006). Marco Barreno et al., 'The Security of Machine Learning', *Machine Learning* 81, no. 2 (2010). Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and Harnessing Adversarial Examples', *arXiv preprint arXiv:1412.6572* (2014). Sandy Huang et al., 'Adversarial Attacks on Neural Network Policies', *arXiv preprint arXiv:1702.02284* (2017). Nicolas Papernot et al., 'Practical Black-Box Attacks against Machine Learning' (paper presented at 'Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security', 2017). Vahid Behzadan and Arslan Munir, 'Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks', *arXiv preprint arXiv:1701.04143* (2017).

program, in that they cause illegitimate data to be classified as legitimate; if this occurs enough, they will threaten the program's usefulness.

A second category of attacks on machine learning programs is known as a causative attack, because hackers attempt to create a weakness that they will later exploit. They will often do this by interfering with the training of the machine learning program through something known as a poisoning attack. In a poisoning attack, the machine learning algorithm is fed carefully-created samples that cause the program to learn the wrong things. For example, a simple attack might feed a spam filter hundreds of messages about cheap imported drugs. The attackers, using a variety of adversarial techniques, can cause the program to assess these messages as legitimate. If the machine then learns that mention of cheap imported drugs is likely to be a sign that a message is benign, future spam messages in that vein are less likely to be blocked. There are numerous examples of these kinds of attacks, especially in programs that continually retrain on recent data; if attackers know that an algorithm is continually retraining itself on incoming data—as opposed to making predictions about data and then discarding the data once it is received—this presents the opportunity to manipulate the learning process.

It took decades after the invention of the computer and the internet for even the basics of cybersecurity to be worked out. Machine learning is comparatively nascent, and the security considerations of machine learning systems' cybersecurity are more nascent still. It suffices here to note that, as machine learning algorithms gain more and more of a role in society, these security concerns must be accorded paramount importance. Especially in cases where the machine will face an intelligent or evolving adversary, security cannot be taken for granted. Security principles must be built in from the start, and operators must be able to adapt flexibly to emerging threats.

Conclusion and Recommendations

We have tried to lay out the enormous opportunities and advances machine learning makes possible. None of these are inevitable. Each requires the groundbreaking work of machine learning researchers—efforts that, like all invention, require as much perspiration as inspiration. But, to make sure the benefits to society are realized, these advances need significant attention from policymakers. More important than any specific policy decision is that policymakers and their staffs begin to understand the key ideas of machine learning and their implications. These issues will only increase in relevance. It is not an exaggeration to say that before long, for example, the majority of Congressional committees will find machine learning impacting the areas they oversee.

We recognize that there is value to particular suggestions. As such, we make seven recommendations for immediate consideration. First, *policymakers should study, or appoint a review group to study, the degree to which machine learning is currently used to reinforce bias in systems to which it is already applied.* The focus on applied systems is vital. The goal here is not to stunt research but to make sure that the technology is not actively causing harm today. Systems like the biased sentencing algorithm mentioned earlier deserve auditing and scrutiny. If machine learning is permitted to reinforce bias, this will stifle its long-run use and value. There is a tradeoff between fairness and accuracy, as mentioned, which must be navigated here. The proper way to do so is context-dependent—no system is perfect—but the tradeoff and its implications should be transparent.

Second, *policymakers should study the ethics and impact of machine learning as it will be used in future applications*. National governments should not rush headlong to develop and deploy machine learning technologies in areas like warfighting without understanding their impact. We recommend significant study, in classified and unclassified environments, of the risks of applying machine learning to these over-the-horizon problems. These studies should examine the risk of bias, of a loss of privacy, and of strategic consequences—for example, how will other nations approach machine learning in their militaries

if the United States continues with the third offset? It will also likely be important to study how the private sector is using or will use machine learning in applied contexts, especially with regards to ethics and impact.

Third, *governments should incentivize the sharing of data within industries and regularly include data openness as a requirement for grants and funding.* Establishing or incentivizing mechanisms for industries, such as healthcare, to establish data interoperability is an essential part of making sure the data can be aggregated and used for analysis. Some cities, such as San Diego, have taken promising first steps by supporting health information exchanges. Furthermore, an enormous amount of data is produced by government-funded studies or organizations. In many of these cases, it would be beneficial for a condition of funding to be the publication of data for secondary research and for use in machine learning applications.

Fourth, *governments at all levels should actively hire individuals to promote the publication of data and the usage of it within the government.* Some cities—again, San Diego is an example—have taken the step of hiring a chief data scientist. The Obama Administration made a big effort to promote open government and make data available. These are positive developments and should serve as models. However, governments should work as well not just to share the data with others, but also to use it to make their own operations more efficient and effective.

Fifth, *governments should expand access to machine learning education and invest in machine learning research*. We believe the technology involved in machine learning is of enormous economic importance. The nations that invest in training the next generation of artificial intelligence scientists will have a significant competitive edge in the global economy. It is essential that socioeconomic factors not limit the opportunity for individuals to work with machine learning; we believe the basics of machine learning can be taught widely, including some foundational ideas in high school. Beyond that, better machine learning will have transformational effects, including ones that save lives. The more individuals who are given the opportunity to understand how machine learning algorithms work, the better—even if those individuals do not end up as research scientists.

Sixth, *governments should particularly encourage the development of privacy- and security-protecting machine learning technologies*. We do not advocate regulation of machine learning programs, however we think it is vital that privacy of data and transparency of algorithms be ensured as much as possible. Governments should particularly encourage and fund research into privacy-protecting machine learning technologies throughout the software and hardware stack. Machine learning will be more trusted in the long run if it is implemented on auditable and tamper-free processing units, using auditable and transparent data mechanisms run by secure algorithms. As the risks of adversarial machine learning become more apparent, this is a vital step, and further investment can build on the impressive work the technology sector has already done.

Seventh, *governments should study the feasibility of explainable machine learning in certain high-priority environments*. As we have discussed, the black box problem is a significant issue in machine learning algorithms with major real-world consequences. Research into explainable machine learning algorithms is ongoing, though virtually all approaches will require tradeoffs of some kind. This sort of research should be encouraged, and policymakers should study the feasibility of explainable machine learning in certain high-priority applications. We do not think an explainability requirement should be applied to every use of machine learning. For example, it may be that a cancer-detecting algorithm does not need to provide explanations, as what is most important is that it detects cancer; on the other hand, an algorithm used for prison sentencing operates in a very different context, where explanations are almost certainly more appropriate. In those contexts, the reasons for decision are an essential part of ensuring not just fairness but the appearance of fairness. Without wide confidence in algorithms and their decisions, machine learning may lose public trust.

To return to where we began: though it may sometimes seem like science fiction, machine learning is not magic. It is neither intrinsically good nor intrinsically evil. It is, perhaps, the ultimate tool. As it becomes more powerful and is applied to more and more problems, the opportunities and dangers of machine learning grow, possibly at an exponential rate. In the end, the promise and peril of machine learning will be determined not by the tool itself, but by those who develop, foster, manage, and wield it. In a democracy, that's all of us.

Works Cited

'3D Makeover for Hyper-Efficient Metalwork', Arup, 11 May 2015, http://www.arup.com/news/2015_05_may/11_may_3d_makeover_for_hyper-efficient_metalwork

Acemoglu, Daron, and Pascual Restrepo, 'The Race between Machine and Man: Implications of Technology for Growth, Factor Shares and Employment', 2016,

———, 'Robots and Jobs: Evidence from US Labor Markets'. (2017).

Alexander, Keith, 'Budget Request for Information Technology and Cyber Operations Programs: Written Testimony of Gen. Keith Alexander', Committee on Armed Services: US House of Representatives, 2012,

Anderson, Nate, '"Anonymized" Data Really Isn't—and Here's Why Not', ArsTechnica, 8 September 2009, https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

Angwin, Julia, Jeff Larson, Surya Mattu, and Laura Kirchner, 'Machine Bias', ProPublica, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

'Apple Previews iOS 10, the Biggest iOS Release Ever', Apple, 13 June 2016, https://www.apple.com/pr/library/2016/06/13Apple-Previews-iOS-10-The-Biggest-iOS-Release-Ever.html

'Autonomous Weapons: An Open Letter from AI & Robotics Researchers', Future of Life Institute, 28 July 2015, https://futureoflife.org/open-letter-autonomous-weapons/

Ayoub, Kareem, and Kenneth Payne, 'Strategy in the Age of Artificial Intelligence'. *Journal of Strategic Studies* 39, no. 5-6 (2016): 793-819.

Banko, Michele, and Eric Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation." Paper presented at 'Proceedings of the 39th Annual Meeting on Association for Computational Linguistics', 2001.

Barreno, Marco, Blaine Nelson, Anthony D Joseph, and JD Tygar, 'The Security of Machine Learning'. *Machine Learning* 81, no. 2 (2010): 121-48.

Barreno, Marco, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. "Can Machine Learning Be Secure?" Paper presented at 'Proceedings of the 2006 ACM Symposium on Information, computer and communications security', 2006.

Behzadan, Vahid, and Arslan Munir, 'Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks'. *arXiv preprint arXiv:1701.04143* (2017).

Brenner, Joel, *Glass Houses*. New York: Penguin, 2014.

Brynjolfsson, Erik, and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company, 2014.

Buchanan, Ben, *The Cybersecurity Dilemma*. New York: Oxford University Press, 2017.

———, 'The Legend of Sophistication in Cyber Operations', Belfer Center for Science and International Affairs, January 2017, https://www.belfercenter.org/sites/default/files/files/publication/Legend Sophistication - web.pdf

———, 'The Life Cycles of Cyber Threats'. *Survival* 58, no. 1 (2016).

Cegłowski, Maciej, 'The Moral Economy of Tech', Society for the Advancement of Socio-Economics, 26 June 2016, http://idlewords.com/talks/sase_panel.htm

Clarke, Arthur C., *Profiles of the Future*. New York: Macmillan, 1973.

Clarke, Richard, and Robert Knake, *Cyberwar*. New York: HarperCollins, 2010.

Cleland, Scott, 'Google's "Infringenovation" Secrets', Forbes, 3 October 2011, https://www.forbes.com/sites/scottcleland/2011/10/03/googles-infringenovation-secrets/ - 539f7bd30a6c

Dwork, Cynthia. "Differential Privacy: A Survey of Results." Paper presented at 'International Conference on Theory and Applications of Models of Computation', 2008.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth, 'The Reusable Holdout: Preserving Validity in Adaptive Data Analysis'. *Science* 349, no. 6248 (2015): 636-38.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through Awareness." Paper presented at 'Proceedings of the 3rd Innovations in Theoretical Computer Science Conference', 2012.

Evans, Richard, and Jim Gao, 'DeepMind AI Reduces Google Data Centre Cooling Bill by 40%', DeepMind, 20 July 2016, https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

Eveleth, Rose, 'Does Crime-Predicting Software Bias Judges? Unfortunately, There's No Data', Motherboard, 18 July 2016, https://motherboard.vice.com/en_us/article/does-crime-predicting-software-bias-judges-unfortunately-theres-no-data

'Fatality Analysis Reporting System', National Highway Traffic Safety Administration: Department of Transportation, 2017, https://www-fars.nhtsa.dot.gov/Main/index.aspx

Fish, Benjamin, Jeremy Kun, and Ádám D. Lelkes. "A Confidence-Based Approach for Balancing Fairness and Accuracy." Paper presented at 'Proceedings of the 2016 SIAM International Conference on Data Mining', 2016.

Francescani, Chris, 'NYPD Report Confirms Manipulation of Crime Stats', Reuters, 9 March 2012, http://www.reuters.com/article/us-crime-newyork-statistics-idUSBRE82818620120309

Freedberg, Sydney, 'Faster Than Thought: DARPA, Artificial Intelligence, & the Third Offset Strategy', Breaking Defense, 11 February 2016, http://breakingdefense.com/2016/02/faster-than-thought-darpa-artificial-intelligence-the-third-offset-strategy/
———, 'War without Fear: DepSecDef Work on How AI Changes Conflict', Breaking Defense, 31 May 2017, http://breakingdefense.com/2017/05/killer-robots-arent-the-problem-its-unpredictable-ai/?utm_source=hs_email&utm_medium=email&utm_content=52548128&_hsenc=p2ANqtz-82jIqhjz8m0NqaiU1c0Xz7Hx1WTT5KExFprW9gni0hN94N3CGCEs88OECYf7JqS21xOBNVH-Veea5eKxYl5Eusnd36lZQ&_hsmi=52548128
Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, *Deep Learning*. MIT Press, 2016.
Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy, 'Explaining and Harnessing Adversarial Examples'. *arXiv preprint arXiv:1412.6572* (2014).
Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans, 'When Will AI Exceed Human Performance? Evidence from AI Experts', ArXiv, 30 May 2017, https://arxiv.org/pdf/1705.08807.pdf
Gunning, David, 'Explainable Artificial Intelligence', Defense Advanced Research Projects Agency, http://www.darpa.mil/program/explainable-artificial-intelligence
Hamburg, Margaret A., and Francis S. Collins, 'The Path to Personalized Medicine'. *New England Journal of Medicine* 363, no. 4 (2010): 301-04.
Hendrickson, Clara, and William A. Galston, 'Automation Presents a Political Challenge, but Also an Opportunity', Brookings Institution, 18 May 2017, https://www.brookings.edu/blog/techtank/2017/05/18/automation-presents-a-political-challenge-but-also-an-opportunity/
Huang, Ling, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. "Adversarial Machine Learning." Paper presented at 'Proceedings of the 4th ACM workshop on Security and artificial intelligence', 2011.
Huang, Sandy, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel, 'Adversarial Attacks on Neural Network Policies'. *arXiv preprint arXiv:1702.02284* (2017).
Johndrow, James E., and Kristian Lum, 'An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction'. *arXiv preprint arXiv:1703.04957* (2017).
Kleinberg, Jon, 'Human Decisions and Machine Predictions'. (2016).
Kumaran, Dharshan, and Demis Hassabis, 'From Pixels to Actions: Human-Level Control through Deep Reinforcement Learning', Google Research, 25 February 2015, https://research.googleblog.com/2015/02/from-pixels-to-actions-human-level.html
Landman, Nathan, Hannah Pang, Eli Ross, and Christopher Williams, 'K-Means Clustering', Brilliant, https://brilliant.org/wiki/k-means-clustering/
Leventhal, Rajiv, 'CMS's Andy Slavitt Says Meaningful Use Will Be Over in 2016', Healthcare Informatics, 12 January 2016, https://www.healthcare-informatics.com/article/cms-s-andy-slavitt-says-mu-will-end-2016-0
Lewis-Kraus, Gideon, 'The Great A.I. Awakening', New York Times, 14 December 2016, https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html?_r=1
Liptak, Adam, 'Sent to Prison by a Software Program's Secret Algorithms', New York Times, 1 May 2017, https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html?_r=0
Liu, Yun, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan*, et al.*, 'Detecting Cancer Metastases on Gigapixel Pathology Images'. *arXiv preprint arXiv:1703.02442* (2017).
Lowry, Stella, and Gordon MacPherson, 'A Blot on the Profession'. *British Medical Journal,* 296, no. 6623 (5 March 1988): 657-58.
Machanavajjhala, Ashwin, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. "Privacy: Theory Meets Practice on the Map." Paper presented at 'IEEE 24th International Conference on Data Engineering', 2008.
Merrett, Rebecca, 'Future of Mobile, IoT Driven by Speech Recognition: Andrew Ng', CIO, 6 May 2015, https://www.cio.com.au/article/574317/future-mobile-iot-driven-by-speech-recognition-andrew-ng/?utm_content=buffer8b8ab&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer
Metz, Cade, 'In Two Moves, AlphaGo and Lee Sedol Redefined the Future', Wired, 16 March 2016, https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/
Miller, Claire Cain, 'Algorithms and Bias: Q. And A. With Cynthia Dwork', New York Times, 10 August 2015, https://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html
———, 'Evidence That Robots Are Winning the Race for American Jobs', New York Times, 28 March 2017, https://mobile.nytimes.com/2017/03/28/upshot/evidence-that-robots-are-winning-the-race-for-american-jobs.html
Mitchell, Tom Michael, *The Discipline of Machine Learning*. Vol. 9: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.
Narayanan, Arvind, and Vitaly Shmatikov. "Robust De-Anonymization of Large Sparse Datasets." Paper presented at 'IEEE Symposium on Security and Privacy', 2008.

'The National Artificial Intelligence Research and Development Strategic Plan', National Science and Technology Council: The White House, 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf

Ng, Andrew, 'AI: The New Electricity', YouTube, 11 June 2016, https://www.youtube.com/watch?v=4eJh-cxfYR4I&feature=youtu.be

———, 'What Artificial Intelligence Can and Can't Do Right Now', Harvard Business Review, 9 November 2016, https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. "Practical Black-Box Attacks against Machine Learning." Paper presented at 'Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security', 2017.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, *et al.*, 'Scikit-Learn: Machine Learning in Python'. *Journal of Machine Learning Research* 12, no. Oct (2011): 2825-30.

Perry, Douglas, 'Sex and Uber's 'Rides of Glory': The Company Tracks Your One-Night Stands -- and Much More', The Oregonian, 20 November 2014, http://www.oregonlive.com/today/index.ssf/2014/11/sex_the_single_girl_and_ubers.html

Pierson, Emma, Sam Corbett-Davies, and Sharad Goel, 'Fast Threshold Tests for Detecting Discrimination'. *arXiv preprint arXiv:1702.08536* (2017).

Porter, Michael E, and Elizabeth Olmsted Teisberg, *Redefining Health Care: Creating Value-Based Competition on Results*. Harvard Business Press, 2006.

'Preparing for the Future of Artificial Intelligence', Executive Office of the President: The White House, 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." Paper presented at 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', 2016.

Rid, Thomas, and Peter McBurney, 'Cyber-Weapons'. *RUSI Journal* 157, no. 1 (2012): 6-13.

Shea, Kristina, Robert Aish, and Marina Gourtovaia, 'Towards Integrated Performance-Driven Generative Design Tools'. *Automation in Construction* 14, no. 2 (2005): 253-64.

Sutton, Richard S, and Andrew G Barto, *Reinforcement Learning: An Introduction*. Vol. 1: MIT press Cambridge, 1998.

Sweeney, Latanya, 'Statement of Latanya Sweeney', Privacy and Integrity Advisory Committee: Department of Homeland Security, 2005, https://www.dhs.gov/xlibrary/assets/privacy/privacy_advcom_06-2005_testimony_sweeney.pdf

Tucker, Patrick, 'The Next Big War Will Turn on AI, Says US Secret-Weapons Czar', Defense One, 28 March 2017, http://www.defenseone.com/technology/2017/03/next-big-war-will-turn-ai-says-pentagons-secret-weapons-czar/136537/

Wang, Dayong, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck, 'Deep Learning for Identifying Metastatic Breast Cancer'. *arXiv preprint arXiv:1606.05718* (2016).

Wang, Rui, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. "Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study." Paper presented at 'Proceedings of the 16th ACM conference on Computer and communications security', 2009.

Warren, Mark, 'The Cure for Cancer Is Data—Mountains of Data', Wired, 19 October 2016, https://www.wired.com/2016/10/eric-schadt-biodata-genomics-medical-research/

Whirl-Carrillo, Michelle, EM McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, RB Altman, and Teri E Klein, 'Pharmacogenomics Knowledge for Personalized Medicine'. *Clinical Pharmacology and Therapeutics* 92, no. 4 (2012): 414.

Work, Robert O., and Shawn Brimley, '20yy: Preparing for War in the Robotic Age', Center for a New American Security, January 2014, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_20YY_WorkBrimley.pdf

The Cyber Security Project
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

www.belfercenter.org/Cyber

Printed in the United States of America