

What AI can and can't do (yet) for your business

Artificial intelligence is a moving target. Here's how to take better aim.

by Michael Chui, James Manyika, and Mehdi Miremadi

Artificial intelligence (AI) seems to be everywhere. We experience it at home and on our phones. Before we know it—if entrepreneurs and business innovators are to be believed—AI will be in just about every product and service we buy and use. In addition, its application to business problem solving is growing in leaps and bounds. And at the same time, concerns about AI's implications are rising: we worry about the impact of AI-enabled automation on the workplace, employment, and society.

A reality sometimes lost amid both the fears and the headline triumphs, such as Alexa, Siri, and AlphaGo, is that the AI technologies themselves—namely, machine learning and its subset, deep learning—have plenty of limitations that will still require considerable effort to overcome. This is an article about those limitations, aimed at helping executives better understand what may be holding back their AI efforts. Along the way, we will also highlight promising advances that are poised to address some of the limitations and create a new wave of opportunities.

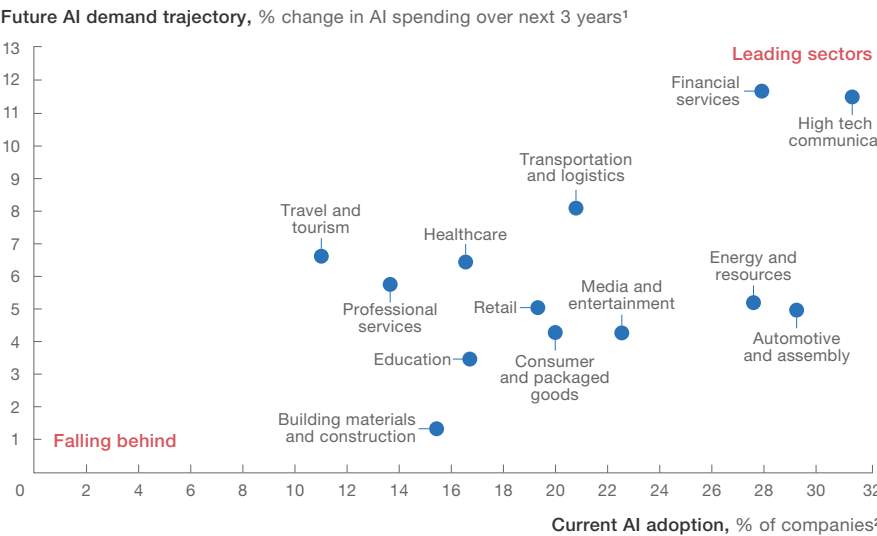
Our perspectives rest on a combination of work at the front lines—researching, analyzing, and assessing hundreds of real-world use cases—and our collaborations with some of the thought leaders, pioneering scientists,

and engineers working at the frontiers of AI. We’ve sought to distill this experience to help executives who often, in our experience, are exposed only to their own initiatives and not well calibrated as to where the frontier is or what the pace setters are already doing with AI.

Simply put, AI’s challenges and limitations are creating a “moving target” problem for leaders: It is hard to reach a leading edge that’s always advancing. It is also disappointing when AI efforts run into real-world barriers, which can lessen the appetite for further investment or encourage a wait-and-see attitude, while others charge ahead. As recent McKinsey Global Institute research indicates, there’s a yawning divide between leaders and laggards in the application of AI both across and within sectors (Exhibit 1).

Executives hoping to narrow the gap must be able to address AI in an informed way. In other words, they need to understand not just where AI can boost innovation, insight, and decision making; lead to revenue growth; and capture of efficiencies—but also where AI *can’t* yet provide value. What’s more, they must appreciate the relationship and distinctions between

Exhibit 1
Leaders in the adoption of AI also intend to **invest more in the near future** compared with laggards.



¹Estimated average, weighted by company size; demand trajectory based on midpoint of range selected by survey respondent.

²Adopting 1 or more AI technologies at scale or in business core; weighted by company size.

Source: McKinsey Global Institute AI adoption and use survey; McKinsey Global Institute analysis

technical constraints and organizational ones, such as cultural barriers; a dearth of personnel capable of building business-ready, AI-powered applications; and the “last mile” challenge of embedding AI in products and processes. If you want to become a leader who understands some of the critical technical challenges slowing AI’s advance and is prepared to exploit promising developments that could overcome those limitations and potentially bend the trajectory of AI—read on.

CHALLENGES, LIMITATIONS, AND OPPORTUNITIES

A useful starting point is to understand recent advances in deep-learning techniques. Arguably the most exciting developments in AI, these advances are delivering jumps in the accuracy of classification and prediction, and are doing so without the usual “feature engineering” associated with traditional supervised learning. Deep learning uses large-scale neural networks that can contain millions of simulated “neurons” structured in layers. The most common networks are called convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These neural networks learn through the use of training data and backpropagation algorithms.

While much progress has been made, more still needs to be done.¹ A critical step is to fit the AI approach to the problem and the availability of data. Since these systems are “trained” rather than programmed, the various processes often require huge amounts of labeled data to perform complex tasks accurately. Obtaining large data sets can be difficult. In some domains, they may simply not be available, but even when available, the labeling efforts can require enormous human resources.

Further, it can be difficult to discern how a mathematical model trained by deep learning arrives at a particular prediction, recommendation, or decision. A black box, even one that does what it’s supposed to, may have limited utility, especially where the predictions or decisions impact society and hold ramifications that can affect individual well-being. In such cases, users sometimes need to know the “whys” behind the workings, such as why an algorithm reached its recommendations—from making factual findings with legal repercussions to arriving at business decisions, such as lending, that have regulatory repercussions—and why certain factors (and not others) were so critical in a given instance.

¹ Stuart Russel et al., “Research priorities for robust and beneficial artificial intelligence,” *AI Magazine*, winter 2015, AAAI.org.

Let's explore five interconnected ways in which these limitations, and the solutions emerging to address them, are starting to play out.

Limitation 1: Data labeling

Most current AI models are trained through “supervised learning.” This means that humans must label and categorize the underlying data, which can be a sizable and error-prone chore. For example, companies developing self-driving-car technologies are hiring hundreds of people to manually annotate hours of video feeds from prototype vehicles to help train these systems. At the same time, promising new techniques are emerging, such as in-stream supervision (demonstrated by Eric Horvitz and his colleagues at Microsoft Research), in which data can be labeled in the course of natural usage.² Unsupervised or semisupervised approaches reduce the need for large, labeled data sets. Two promising techniques are reinforcement learning and generative adversarial networks.

Reinforcement learning. This unsupervised technique allows algorithms to learn tasks simply by trial and error. The methodology hearkens to a “carrot and stick” approach: for every attempt an algorithm makes at performing a task, it receives a “reward” (such as a higher score) if the behavior is successful or a “punishment” if it isn't. With repetition, performance improves, in many cases surpassing human capabilities—so long as the learning environment is representative of the real world.

Reinforcement learning has famously been used in training computers to play games—most recently, in conjunction with deep-learning techniques. In May 2017, for example, it helped the AI system AlphaGo to defeat world champion Ke Jie in the game of Go. In another example, Microsoft has fielded decision services that draw on reinforcement learning and adapt to user preferences. The potential application of reinforcement learning cuts across many business arenas. Possibilities include an AI-driven trading portfolio that acquires or loses points for gains or losses in value, respectively; a product-recommendation engine that receives points for every recommendation-driven sale; and truck-routing software that receives a reward for on-time deliveries or reducing fuel consumption.

Reinforcement learning can also help AI transcend the natural and social limitations of human labeling by developing previously unimagined solutions and strategies that even seasoned practitioners might never have

² Eric Horvitz, “Machine learning, reasoning, and intelligence in daily life: Directions and challenges,” *Proceedings of Artificial Intelligence Techniques for Ambient Intelligence*, Hyderabad, India, January 2007.

considered. Recently, for example, the system AlphaGo Zero, using a novel form of reinforcement learning, defeated its predecessor AlphaGo after learning to play Go from scratch. That meant starting with completely random play against itself rather than training on Go games played by and with humans.³

Generative adversarial networks (GANs). In this semisupervised learning method, two networks compete against each other to improve and refine their understanding of a concept. To recognize what birds look like, for example, one network attempts to distinguish between genuine and fake images of birds, and its opposing network attempts to trick it by producing what look very much like images of birds, but aren't. As the two networks square off, each model's representation of a bird becomes more accurate.

The ability of GANs to generate increasingly believable examples of data can significantly reduce the need for data sets labeled by humans. Training an algorithm to identify different types of tumors from medical images, for example, would typically require millions of human-labeled images with the type or stage of a given tumor. By using a GAN trained to generate increasingly realistic images of different types of tumors, researchers could train a tumor-detection algorithm that combines a much smaller human-labeled data set with the GAN's output.

While the application of GANs in precise disease diagnoses is still a way off, researchers have begun using GANs in increasingly sophisticated contexts. These include understanding and producing artwork in the style of a particular artist and using satellite imagery, along with an understanding of geographical features, to create up-to-date maps of rapidly developing areas.

Limitation 2: Obtaining massive training data sets

It has already been shown that simple AI techniques using linear models can, in some cases, approximate the power of experts in medicine and other fields.⁴ The current wave of machine learning, however, requires training data sets that are not only labeled but also sufficiently large and comprehensive. Deep-learning methods call for thousands of data records for models to become relatively good at classification tasks and, in some cases, millions for them to perform at the level of humans.⁵

³ Demis Hassabis et al., *AlphaGo Zero: Learning from scratch*, deepmind.com.

⁴ Robyn M. Dawes, "The robust beauty of improper linear models in decision making," *American Psychologist*, 1979, Volume 34, Number 7, pp. 571–82.

⁵ Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016.

The complication is that massive data sets can be difficult to obtain or create for many business use cases (think: limited clinical-trial data to predict treatment outcomes more accurately). And each minor variation in an assigned task could require another large data set to conduct even more training. For example, teaching an autonomous vehicle to navigate a mining site where the weather continually changes will require a data set that encompasses the different environmental conditions the vehicle might encounter.

One-shot learning is a technique that could reduce the need for large data sets, allowing an AI model to learn about a subject when it's given a small number of real-world demonstrations or examples (even one, in some cases). AI's capabilities will move closer to those of humans, who can recognize multiple instances of a category relatively accurately after having been shown just a single sample—for example, of a pickup truck. In this still-developing methodology, data scientists would first pre-train a model in a simulated virtual environment that presents variants of a task or, in the case of image recognition, of what an object looks like. Then, after being shown just a few real-world variations that the AI model did *not* see in virtual training, the model would draw on its knowledge to reach the right solution.⁶

This sort of one-shot learning could eventually help power a system to scan texts for copyright violations or to identify a corporate logo in a video after being shown just one labeled example. Today, such applications are only in their early stages. But their utility and efficiency may well expand the use of AI quickly, across multiple industries.

Limitation 3: The explainability problem

Explainability is not a new issue for AI systems.⁷ But it has grown along with the success and adoption of deep learning, which has given rise both to more diverse and advanced applications and to more opaqueness. Larger and more complex models make it hard to explain, in human terms, why a certain decision was reached (and even harder when it was reached in real time). This is one reason that adoption of some AI tools remains low in application areas where explainability is useful or indeed required. Furthermore, as the application of AI expands, regulatory requirements could also drive the need for more explainable AI models.⁸

⁶ Yan Duan et al., *One-shot imitation learning*, December 2017, [arxiv.org](https://arxiv.org/abs/1704.07462).

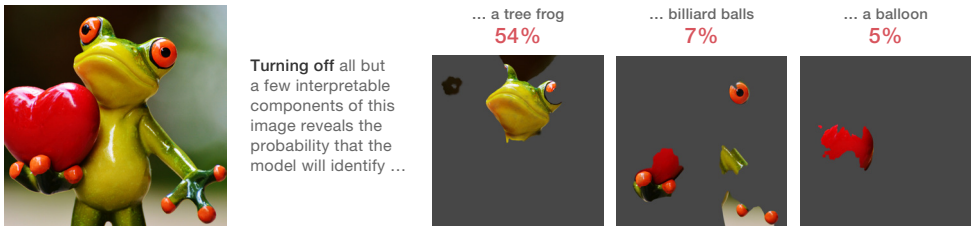
⁷ Eric Horvitz et al., "The use of a heuristic problem-solving hierarchy to facilitate the explanation of hypothesis-directed reasoning," *Proceedings of Medinfo*, October 1986, pp. 27–31.

⁸ See, for example, the European Union's proposed General Data Protection Regulation, which would introduce new requirements for the use of data.

Exhibit 2

New techniques hold promise for making AI more transparent.

LIME¹ is a sensitivity analysis that reveals which parts of an input matter most to the eventual output.



Attention shines a spotlight on where the model is looking when it makes a particular decision.

Words relevant to food quality ...

The fajita we tried was **tasteless** and **burned** and the **mole sauce** was **way too sweet**.

... or to service

They have one of the **fastest delivery times** in the **city**.

¹LIME = local-interpretable-model-agnostic explanations.
Source: Carlos Guestrin, Marco Tulio Ribeiro, and Sameer Singh, "Introduction to local interpretable model-agnostic explanations (LIME)," August 12, 2016, O'Reilly, oreilly.com; Minlie Huang, Yequan Wang, Li Zhao, and Xiaoyan Zhu, *Attention-based LSTM for aspect-level sentiment classification*, Tsinghua University; Pixabay

Two nascent approaches that hold promise for increasing model transparency are local-interpretable-model-agnostic explanations (LIME) and attention techniques (Exhibit 2). LIME attempts to identify which parts of input data a trained model relies on most to make predictions in developing a proxy interpretable model. This technique considers certain segments of data at a time and observes the resulting changes in prediction to fine-tune the proxy model and develop a more refined interpretation (for example, by excluding eyes rather than, say, noses to test which are more important for facial recognition). Attention techniques visualize those pieces of input data that a model considers most as it makes a particular decision (such as focusing on a mouth to determine if an image depicts a human being).

Another technique that has been used for some time is the application of generalized additive models (GAMs). By using single-feature models, GAMs limit interactions between features, thereby making each one more easily

interpretable by users.⁹ Employing these techniques, among others, to demystify AI decisions is expected to go a long way toward increasing the adoption of AI.

Limitation 4: Generalizability of learning

Unlike the way humans learn, AI models have difficulty carrying their experiences from one set of circumstances to another. In effect, whatever a model has achieved for a given use case remains applicable to that use case only. As a result, companies must repeatedly commit resources to train yet another model, even when the use cases are very similar.

One promising response to this challenge is transfer learning.¹⁰ In this approach, an AI model is trained to accomplish a certain task and then quickly applies that learning to a similar but distinct activity. DeepMind researchers have also shown promising results with transfer learning in experiments in which training done in simulation is then transferred to real robotic arms.¹¹

As transfer learning and other generalized approaches mature, they could help organizations build new applications more quickly and imbue existing applications with more diverse functionality. In creating a virtual personal assistant, for example, transfer learning could generalize user preferences in one area (such as music) to others (books). And users are not restricted to digital natives. Transfer learning can enable an oil-and-gas producer, for instance, to expand its use of AI algorithms trained to provide predictive maintenance for wells to other equipment, such as pipelines and drilling platforms. Transfer learning even has the potential to revolutionize business intelligence: consider a data-analyzing AI tool that understands how to optimize airline revenues and can then adapt its model to changes in weather or local economics.

Another approach is the use of something approximating a generalized structure that can be applied in multiple problems. DeepMind's AlphaZero, for example, has made use of the same structure for three different games: it has been possible to train a new model with that generalized structure

⁹ Yin Lou, Rich Caruana, and Johannes Gehrke, "Intelligible models for classification and regression," *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, 2012, pp. 150–58.

¹⁰ For an earlier example application, see John Guttag, Eric Horvitz, and Jenna Wiens, "A study in transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions," *Journal of the American Medical Informatics Association*, 2014, Volume 21, Number 4, pp. 699–706.

¹¹ Andrei A. Rusu et al., *Sim-to-real robot learning from pixels with progressive nets*, October 2016, arxiv.org.

to learn chess in a single day, and it then soundly beat a world-champion chess program.¹²

Finally, consider the possibilities in emerging meta-learning techniques that attempt to automate the design of machine-learning models. The Google Brain team, for example, uses AutoML to automate the design of neural networks for classifying images in large-scale data sets. These techniques now perform as well as those designed by humans.¹³ That's a promising development, particularly as talent continues to be in short supply for many organizations. It's also possible that meta-learning approaches will surpass human capabilities and yield even better results. Importantly, however, these techniques are still in their early days.

Limitation 5: Bias in data and algorithms

So far, we've focused on limitations that could be overcome through technical solutions already in the works, some of which we have described. Bias is a different kind of challenge. Potentially devastating social repercussions can arise when human predilections (conscious or unaware) are brought to bear in choosing which data points to use and which to disregard. Furthermore, when the process and frequency of data collection itself are uneven across groups and observed behaviors, it's easy for problems to arise in how algorithms analyze that data, learn, and make predictions.¹⁴ Negative consequences can include misinformed recruiting decisions, misrepresented scientific or medical prognoses, distorted financial models and criminal-justice decisions, and misapplied (virtual) fingers on legal scales.¹⁵ In many cases, these biases go unrecognized or disregarded under the veil of "advanced data sciences," "proprietary data and algorithms," or "objective analysis."

As we deploy machine learning and AI algorithms in new areas, there probably will be more instances in which these issues of potential bias become baked into data sets and algorithms. Such biases have a tendency to stay embedded because recognizing them, and taking steps to address them, requires a deep mastery of data-science techniques, as well as a more meta-understanding of existing social forces, including data collection. In all,

¹² David Silver et al., *Mastering chess and shogi by self-play with a general reinforcement learning algorithm*, December 2017, [arxiv.org](https://arxiv.org/abs/1712.04801).

¹³ *Google Research Blog*, "AutoML for large scale image classification and object detection," blog entry by Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc Le, November 2, 2017, [research.googleblog.com](https://research.googleblog.com/2017/11/02/automl-for-large-scale-image-classification-and-object-detection.html).

¹⁴ Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, *Inherent trade-offs in the fair determination of risk scores*, November 2016, [arxiv.org](https://arxiv.org/abs/1611.08226).

¹⁵ See the work of Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, and Terry Parris Jr. of ProPublica.

debiasing is proving to be among the most daunting obstacles, and certainly the most socially fraught, to date.

There are now multiple research efforts under way, as well as efforts to capture best practices, that address these issues in academic, nonprofit, and private-sector research. It's none too soon, because the challenge is likely to become even more critical, and more questions will arise. Consider, for example, the fact that many of these learning and statistically based predictive approaches implicitly assume that the future will be like the past. What should we do in sociocultural settings where efforts are under way to spur change—and where making decisions based on past behavior could inhibit progress (or, worse, build in resistance to change)? A wide variety of leaders, including business leaders, may soon be called upon to answer such questions.

HITTING THE MOVING TARGET

Solutions to the limitations we have described, along with the widespread commercial implementation of many of the advances described here, could be years away. But the breathtaking range of possibilities from AI adoption suggests that the greatest constraint for AI may be imagination. Here are a few suggestions for leaders striving to stay ahead of—or at least not fall too far behind—the curve:

Do your homework, get calibrated, and keep up. While most executives won't need to know the difference between convolutional and recurrent neural networks, you should have a general familiarity with the capabilities of today's tools, a sense of where short-term advances are likely to occur, and a perspective on what's further beyond the horizon. Tap your data-science and machine-learning experts for their knowledge, talk to some AI pioneers to get calibrated, and attend an AI conference or two to help you get the real facts; news outlets can be helpful, but they can also be part of the hype machine. Ongoing tracking studies by knowledgeable practitioners, such as the AI Index (a project of the Stanford-based One Hundred Year Study on Artificial Intelligence), are another helpful way to keep up.¹⁶


Adopt a sophisticated data strategy. AI algorithms need assistance to unlock the valuable insights lurking in the data your systems generate. You can help by developing a comprehensive data strategy that focuses not only on the technology required to pool data from disparate systems but also on data availability and acquisition, data labeling, and data governance. Although newer techniques promise to reduce the amount of data required for training

¹⁶ See the AI Index (aiindex.org) and the One Hundred Year Study (ai100.stanford.edu).

AI algorithms, data-hungry supervised learning remains the most prevalent technique today. And even techniques that aim to minimize the amount of data required still need *some* data. So a key part of this is fully knowing your own data points and how to leverage them.

Think laterally. Transfer-learning techniques remain in their infancy, but there are ways to leverage an AI solution in more than one area. If you solve a problem such as predictive maintenance for large warehouse equipment, can you also apply the same solution to consumer products? Can an effective next-product-to-buy solution be used in more than one distribution channel? Encourage business units to share knowledge that may reveal ways to use your best AI solutions and thinking in more than one area of the company.

Be a trailblazer. Keeping up with today's AI technologies and use cases is not enough to remain competitive for the long haul. Engage your data-science staff or partner with outside experts to solve a high-impact use case with nascent techniques, such as the ones discussed in this article, that are poised for a breakthrough. Further, stay informed about what's possible and what's available. Many machine-learning tools, data sets, and trained models for standard applications (including speech, vision, and emotion detection) are being made widely available. Sometimes they come in open source and in other cases through application programming interfaces (APIs) created by pioneering researchers and companies. Keep an eye on such possibilities to boost your odds of staking out a first-mover or early-adopter advantage.

The promise of AI is immense, and the technologies, tools, and processes needed to fulfill that promise haven't fully arrived. If you think you can let the technology develop and then be a successful fast follower, think again. It's very difficult to leapfrog from a standing start, particularly when the target is moving so rapidly and you don't understand what AI tools can and can't do now. With researchers and AI pioneers poised to solve some of today's thorniest problems, it's time to start understanding what is happening at the AI frontier so you can position your organization to learn, exploit, and maybe even advance the new possibilities. 

Michael Chui is a partner of the McKinsey Global Institute (MGI) and is based in McKinsey's San Francisco office; **James Manyika** is the chairman of MGI and a senior partner in the San Francisco office; and **Mehdi Miremadi** is a partner in the Chicago office.

The authors wish to thank Jack Clark at OpenAI, Jeffrey Dean at Google Brain, Professor Barbara Grosz at Harvard University, Demis Hassabis at DeepMind, Eric Horvitz at Microsoft Research, and Martin Wicke for their insights on the ideas in this article. They also wish to thank their McKinsey colleagues Steven Adler, Ali Akhtar, Adib Ayay, Ira Chadha, Rita Chung, Nicolaus Henke, Sankalp Malhotra, and Pieter Nel for their contributions to this article.