McKinsey&Company

The real-world potential and limitations of artificial intelligence

McKinsey Quarterly April 2018

Artificial intelligence has the potential to create trillions of dollars of value across the economy—if business leaders work to understand what AI can and cannot do.

In this episode of the *McKinsey Podcast,* McKinsey Global Institute partner Michael Chui and MGI chairman and director James Manyika speak with McKinsey Publishing's David Schwartz about the cutting edge of artificial intelligence.

Podcast transcript

David Schwartz: Hello, and welcome to the *McKinsey Podcast*. I'm David Schwartz with McKinsey Publishing. Today, we're going to be journeying to the frontiers of artificial intelligence. We'll touch on what AI's impact could be across multiple industries and functions. We'll also explore limitations that, at least for now, stand in the way.

I'm joined by two McKinsey leaders who are at the point of the spear, Michael Chui, based in San Francisco and a partner with the McKinsey Global Institute, and James Manyika, the chairman of the McKinsey Global Institute and a senior partner in our San Francisco office. Michael and James, welcome.

James Manyika: Thanks for having us.

Michael Chui: Great to be here.

David Schwartz: Michael, where do we see the most potential from AI?

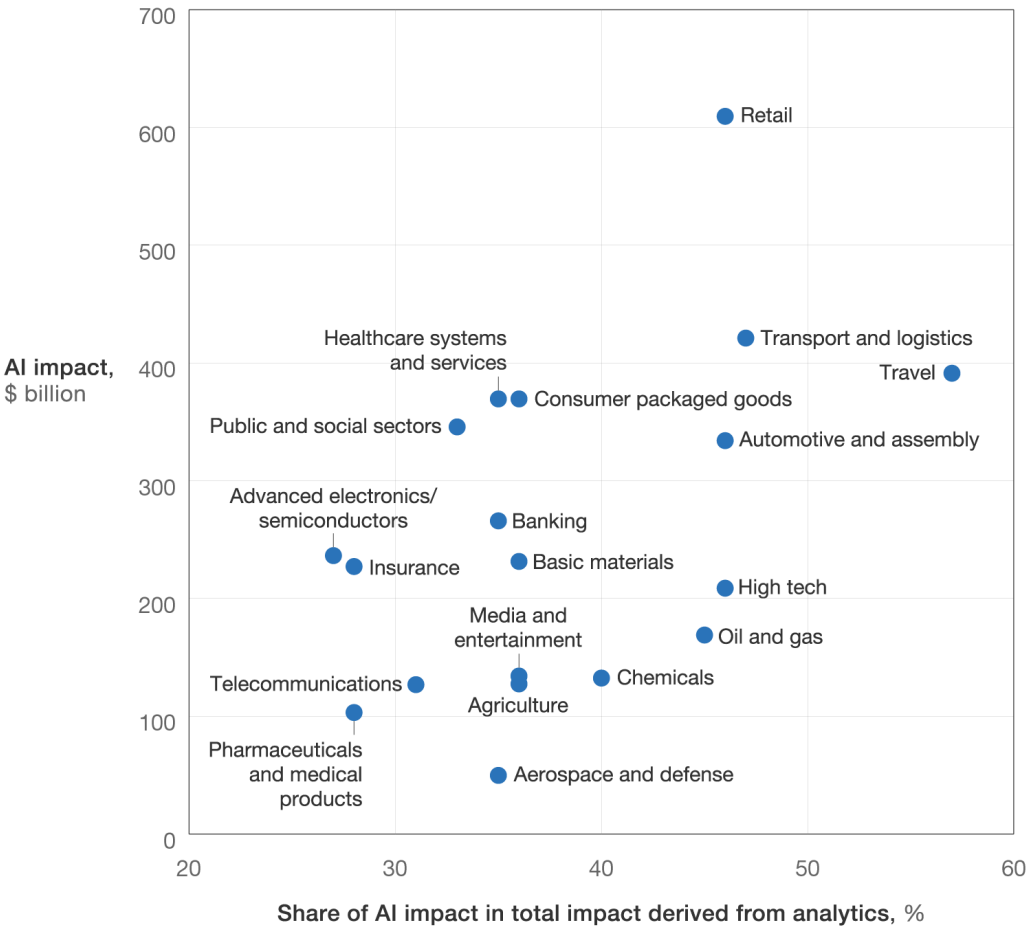
Michael Chui: The number-one thing that we know is just the widespread potential applicability. That said, we're quite early in terms of the adoption of these technologies, so there's a lot of runway to go. One of the other things that we've discovered is that one way to think about where the potential for AI is, is just follow the money.

If you're a company where marketing and sales is what drives the value, that's actually where AI can create the most value. If you're a company where operational excellence matters the most to you, that's where you can create the most value with AI. If you're an insurance company, or

if you're a bank, then risk is really important to you, and that's another place where AI can add value. It goes through everything from managing human capital and analyzing your people's performance and recruitment, et cetera, all through the entire business system. We see the potential for trillions of dollars of value to be created annually across the entire economy [Exhibit 1].

Exhibit 1

Artificial intelligence (AI) has the potential to create value across sectors.



McKinsey&Company | **Source:** McKinsey Global Institute analysis

David Schwartz: Well, it certainly sounds like there's a lot of potential and a lot of value yet to be unleashed. James, can you come at it from the other direction? What are the big limitations of AI today? And what do these mean in practical terms for business leaders?

James Manyika: When we think about the limitations of AI, we have to keep in mind that this is still a very rapidly evolving set of techniques and technologies, so the science itself and the techniques themselves are still going through development.

When you think about the limitations, I would think of them in several ways. There are limitations that are purely technical. Questions like, can we actually explain what the algorithm is doing? Can we interpret why it's making the choices and the outcomes and predictions that it's making? Then you've also got a set of practical limitations. Questions like, is the data actually available? Is it labeled? We'll get into that in a little bit.

But I'd also add a third limitation. These are limitations that you might call limitations in use. These are what leads you to questions around, how transparent are the algorithms? Is there any bias in the data? Is there any bias in the way the data was collected?

David Schwartz: Michael, let's drill down on a first key limitation, data labeling. Can you describe the challenge and some possible ways forward?

Michael Chui: One of the things that's a little bit new about the current generations of AI is what we call machine learning—in the sense that we're not just programming computers, but we're training them; we're teaching them.

The way we train them is to give them this labeled data. If you're trying to teach a computer to recognize an object within an image, or if you're trying to teach your computer to recognize an anomaly within a data stream that says a piece of machinery is about to break down, the way you do that is to have a bunch of labeled data and say, "Look, in these types of images, the object is present. In these types of images, the object's not present. In these types of data streams, the machine's about to break, and in these types of data streams, the machine's not about to break."

We have this idea that machines will train themselves. Actually, we've generated a huge amount of work for people to do. Take, for example, self-driving cars. These self-driving cars have cameras on them, and one of the things that they're trying to do is collect a bunch of data by driving around.

It turns out, there is an army of people who are taking the video inputs from this data and then just tracing out where the other cars are—where the lane markers are as well. So, the funny thing is, we talk about these AI systems automating what people do. In fact, it's generating a whole bunch of manual labor for people to do.

James Manyika: I know this large public museum where they get students to literally label pieces of art—that's a cat, that's a dog, that's a tree, that's a shadow. They just label these different pieces of art so that algorithms can then better understand them and be able to make predictions.

In older versions of this, people were identifying cats and dogs. There have been teams, for example, in the UK that were going to identify different breeds of dogs for the purposes of labeling data images for dogs so that when algorithms use that data, they know what it is. The same thing is happening in a lot of medical applications, where people have been labeling different kinds of tumors, for example, so that when machines read those images, they can

better understand what's a tumor and what kind of tumor is it. But it has taken people to label those different tumors for that to then be useful for the machines.

Michael Chui: A medical diagnosis is the perfect example. So, for this idea of having a system that looks at X-rays and decides whether or not people have pneumonia, you need the data to tell whether or not this X-ray was associated with somebody who had pneumonia or didn't have pneumonia. Collecting that data is an incredibly important thing, but labeling it is absolutely necessary.

David Schwartz: Let's talk about ways to possibly solve it. I know that there are two techniques in supervised learning that we're hearing a lot about. One is reinforcement learning, and the other is GANs [generative adversarial networks]. Could you speak about those?

> "Companies and organizations that are taking AI seriously are playing these multiyear games to acquire the data that they need."

Michael Chui: A number of these techniques are meant to basically create more examples that allow you to teach the machine, or have it learn.

Reinforcement learning has been used to train robots, in the sense that if the robot does the behavior that you want it to, you reward the robot for doing it. If it does a behavior you don't want it to do, you give it negative reinforcement. In that case, what you have is a function that says whether you did something good or bad. Rather than having a huge set of labeled data, you just have a function that says you did good or you did the wrong thing. That's one way to get around label data—by having a function that tells you whether you did the right thing.

With GANs, which stands for generative adversarial networks, you basically have two networks, one that's trying to generate the right thing; the other one is trying to discriminate whether you're generating the right thing. Again, it's another way to get around one potential limitation of having huge amounts of label data in the sense that you have two systems that are competing against each other in an adversarial way. It's been used for doing all kinds of things. The generative—the "G" part of it—is what's remarkable. You can generate art in the style of another artist. You can generate architecture in the style of other things that you've observed. You can generate designs that look like other things that you might have observed before.

James Manyika: The one thing I would add about GANs is that, in many respects, they're a form of semisupervised learning techniques in the sense that they typically start with some initial labeling but then, in a generative way, build on it— in this adversarial, kind of a contest way.

There's also a whole host of other techniques that people are experimenting with. One of the things, for example, is researchers at Microsoft Research Lab have been working on instream labeling, where you'll actually label the data through use. You're trying to interpret

based on how the data's being used, what it actually means. This idea of instream labeling has been around for quite a while, but in recent years, it has started to demonstrate some quite remarkable results. This problem of labeling is one we're going to be with for quite a while.

David Schwartz: What about limitations when there is not enough data?

Michael Chui: One of the things that we've heard from Andrew Ng, who's one of the leaders in machine learning and AI, is that companies and organizations that are taking AI seriously are playing these multiyear games to acquire the data that they need.

In the physical world, whether you're doing self-driving cars or drones, it takes time to go out and drive a whole bunch of streets or fly a whole bunch of things. To try to improve the speed at which you can learn some of those things, one of the things you can do is simulate environments. By creating these virtual environments—basically within a data center, basically within a computer—you can run a whole bunch more trials and learn a whole bunch more things through simulation. So, when you actually end up in the physical world, you've come to the physical world with your AI already having learned a bunch of things in simulation.

"That's the holy-grail question: How do you build generalizable systems that can learn anything?"

James Manyika: A good example of that is some of the demonstrations, for example, that the team at DeepMind Technologies has done. They've done a lot of simulated training for robotic arms, where much of the manipulation techniques that these robotic arms have been able to develop and learn was from having actually been done in simulation—way before the robot arm was even applied to the real world. When it shows up in the real world, it comes with these prelearned data sets that have come out of simulation as a way to get around the limitations of data.

David Schwartz: It sounds like we may be considering a deeper issue—what machine intelligence actually means. How can we move from a process of rote inputs and set outputs to something more along the lines of the ways that humans learn?

James Manyika: That's, in some ways, the holy-grail question, which is: How do you build generalizable systems that can learn anything? Humans are remarkable in the sense that we can take things we've learned over here and apply them to totally different problems that we may be seeing for the first time. This has led to one big area of research that's typically referred to as transfer learning, the idea of, how do you take models or learnings or insights from one arena and apply them to another? While we're making progress in transfer learning, it's actually one of the harder problems to solve. And there, you're finding new techniques.

This idea of simulating learning where you generate data sets and simulations is one way to do that. AlphaGo Zero, which is a more interesting version, if you like, of AlphaGo, has learned to play three different games but has just a generalized structure of games. Through that, it's

been able to learn chess and Go—by having a generalized structure. But even that is limited in the sense that it's still limited to games that take a certain form.

Michael Chui: In the AI field, what we're relearning, which neurologists have known for a long time, is that as people, we don't come as tabula rasa. We actually have a number of structures in our brain that are optimized for certain things, whether it's understanding language or behavior, physical behavior, et cetera. People like Geoff Hinton are using capsules and other types of concepts. This idea of embedding some learning in the structure of the systems that we're using is something that we've seen as well. And so, you wonder whether for transfer learning, part of the solution is understanding that we don't start from nothing. We start from systems that have some configuration already, and that helps us be able to take certain learnings from one place to another because, actually, we're set up to do that.

James Manyika: In fact, Steve Wozniak has come out with certain suggestions, and this has led to all kinds of questions about what's the right Turing test or the kind of test you can come up with generalized learning. One version that he has is the so-called "coffee test," which is, the day we can get a system that could walk into an unknown American household and make a cup of coffee. That's pretty remarkable, because that requires being able to interpret a totally unknown environment, being able to discover things in a totally unknown place, and being able to make something with unknown equipment in a particular household.

There are a lot of general problems that need to be solved along the way of making a cup of coffee in an unknown household, which may sound trivial compared to solving very narrow, highly technical, specific problems which we think of as remarkable. The more we can then look to solving what are generalized often as, quite frankly, garden-variety, real-world problems, those might actually be the true tests of whether we have generalized systems or not.

And it is important to remember, by the way, as we think about all the exciting stuff that's going on in AI and machine learning, that the vast majority—whether it's the techniques or even the applications—are mostly solving very specific things. They're solving natural-language processing; they're solving image recognition; they're doing very, very specific things. There's a huge flourishing of that, whereas the work going toward solving the more generalized problems, while it's making progress, is proceeding much, much more slowly. We shouldn't confuse the progress we're making on these more narrow, specific problem sets to mean, therefore, we have created a generalized system.

There's another limitation, which we should probably discuss, David—and it's an important one for lots of reasons. This is the question of "explainability." Essentially, neural networks, by their structure, are such that it's very hard to pinpoint why a particular outcome is what it is and where exactly in the structure of it something led to a particular outcome.

David Schwartz: Right. I'm hearing that we're dealing with very complicated problems, very complex issues. How would someone, outside in, ever understand what may appear to be—may in fact be—almost a black box?

James Manyika: This is the question of explainability, which is: How do we even know that? You think about where we start applying these systems in the financial world—for example, to lending. If we deny you for a mortgage application, you may want to know why. What is the data point or feature set that led to that decision? If you apply the system set to the criminal-justice system, if somebody's been let out on bail and somebody else wasn't, you may want to understand why it is that we came to that conclusion. It may also be an important question for purely research purposes, where you're trying to self-discover particular behaviors, and so you're trying to understand what particular part of the data leads to a particular set of behaviors.

This is a very hard problem structurally. The good news, though, is that we're starting to make progress on some of these things. One of the ways in which we're making progress is with so-called GANs. These are more generalized, additive models where, as opposed to taking massive amounts of models at the same time, you almost take one feature model set at a time, and you build on it.

For example, when you apply the neural network, you're exploring one particular feature, and then you layer on another feature; so, you can see how the results are changing based on this kind of layering, if you like, of different feature models. You can see, when the results shift, which model feature set seemed to have made the biggest difference. This is a way to start to get some insight into what exactly is driving the behaviors and outcomes you're getting.

Michael Chui: One of the other big drivers for explainability is regulation and regulators. If a car decides to make a left turn versus a right turn, and there's some liability associated with that, the legal system will want to ask the question, "Why did the car make the left turn or the right turn?" In the European Union, there's the General Data Protection Regulation that will require explainability for certain types of decisions that these machines might make. The machines are completely deterministic. You could say, "Here are a million weights that are associated with our simulated neurons. Here's why." But that's not engaging to a human being.

Another technique is an acronym, LIME, which is locally interpretable model-agnostic explanations. The idea there is from the outside in—rather than look at the structure of the model, just be able to perturb certain parts of the model and the inputs and see whether that makes a difference on the outputs. If you're taking a look at an image and trying to recognize whether an object is a pickup truck or an ordinary sedan, you might say, "If I change the wind screen on the inputs, does that cause me to have a different output? On the other hand, if I change the back end of the vehicle, it looks like that makes a difference." That says, that what this model is paying attention to as it's determining whether it's a sedan or a pickup truck is the back part of the vehicle. It's basically doing experiments on the model in order to figure out what makes a difference. Those are some of the techniques that people are trying to use in order to explain how these systems work.

David Schwartz: At some level, I'm hearing from the questions and from what the rejoinder might be that there's a very human element. A question would be: Why is the answer such and such? And the answer could be, it's the algorithm. But somebody built that algorithm, or

somebody—or a team of somebodies—and machines built that algorithm. That brings us to a limitation that is not quite like the others: bias—human predilections. Could you speak a little bit more about what we're up against, James?

> "It becomes very, very important to think through what might be the inherent biases in the data, in any direction—either in the actual way it's constructed, or even the way it's collected, or the degree of sampling of the data and the granularity of it."

James Manyika: The question of bias is a very important one. And I'd put it into two parts.

Clearly, these algorithms are, in some ways, a big improvement on human biases. This is the positive side of the bias conversation. We know that, for example, sometimes, when humans are interpreting data on CVs [curriculum vitae], they might gravitate to one set of attributes and ignore some other attributes because of whatever predilections that they bring. There's a big part of this in which the application of these algorithms is, in fact, a significant improvement compared to human biases. In that sense, this is a good thing. We want those kinds of benefits.

But I think it's worth having the second part of the conversation, which is, even when we are applying these algorithms, we do know that they are creatures of the data and the inputs you put in. If those inputs you put in have some inherent biases themselves, you may be introducing different kinds of biases at much larger scale.

The work of people like Julia Angwin and others has actually shown this if the data collected is already biased. If you take policing as an example, we know that there are some communities that are more heavily policed. There's a much larger police presence. Therefore, the data we've got and that's collected about those environments is much, much, much higher. If we then start to compare, say, two neighborhoods, one where it's oversampled—meaning there's lots and lots of data available for it because there's a larger police presence—versus another one where there isn't much policing so, therefore, there isn't much data available, we may draw the wrong conclusions about the heavily policed observed environment, just simply because there's more data available for it versus the other one.

The biases can go another way. For example, in the case of lending, the implications might go the other way. For populations or segments where we have lots and lots of financial data about them, we may actually make good decisions because the data is largely available, versus in another environment where we're talking about a segment of the population we don't know much about, and the little bit that we know sends the decision off in one way. And so, that's another example where the undersampling creates a bias.

The point about this second part is that I think it becomes very, very important to make sure that we think through what might be the inherent biases in the data, in any direction, that might be in the data set itself—either in the actual way it's constructed, or even the way it's collected, or the degree of sampling of the data and the granularity of it. Can we debias that in some fundamental way?

This is why the question of bias, for leaders, is particularly important, because it runs a risk of opening companies up to all kinds of potential litigation and social concern, particularly when you get to using these algorithms in ways that have social implications. Again, lending is a good example. Criminal justice is another example. Provision of healthcare is another example. These become very, very important arenas to think about these questions of bias.

Michael Chui: Some of the difficult cases where there's bias in the data, at least in the first instance, isn't around, as a primary factor, people's inherent biases about choosing either one or the other. It is around, in many cases, these ideas about sampling—sampling bias, data-collection bias, et cetera—which, again, is not necessarily about unconscious human bias but an artifact of where the data came from.

There's a very famous case, less AI related, where an American city used an app in the early days of smartphones that determined where potholes were based on the accelerometer shaking when you drove over a pothole. Strangely, it discovered that if you looked at the data, it seemed that there were more potholes in affluent parts of the city. That had nothing to do with the fact there were actually more potholes in that part of the city, but you had more signals from that part of the city because more affluent people had more smartphones at the time. That's one of those cases where it wasn't because of any intention to not pay attention to certain parts of the city. Understanding the providence of data—understanding what's being sampled—is incredibly important.

There's another researcher who has a famous TED Talk, Joy Buolamwini at MIT Media Lab. She does a lot of work on facial recognition, and she's a black woman. And she says, "Look, a lot of the other researchers are more male and more pale than I am. And as a result, the accuracy for certain populations in facial recognition is far higher than it is for me." So again, it's not necessarily because people are trying to exclude populations, although sometimes that happens, it really has to do with understanding the representativeness of the sample that you're using in order to train your systems.

So, as a business leader, you need to understand, if you're going to train machine-learning systems: How representative are the training sets there that you're using?

"People forget that one of the things in the AI machine-deep-learning world is that many researchers are using largely the same data sets that are shared—that are public."

James Manyika: It actually creates an interesting tension. That's why I described the part one and the part two. Because in the first instance, when you look at the part-one problem, which is the inherent human biases in normal day-to-day hiring and similar decisions, you get very excited about using AI techniques. You say, "Wow, for the first time, we have a way to get past these human biases in everyday decisions." But at the same time, we should be thoughtful about where that takes us to when you get to these part-two problems, where you now are using large data sets that have inherent biases.

I think people forget that one of the things in the AI machine-deep-learning world is that many researchers are using largely the same data sets that are shared—that are public. Unless you happen to be a company that has these large, proprietary data sets, people are using this famous CIFAR data set, which is often used for object recognition. It's publicly available. Most people benchmark their performance on image recognition based on these publicly available data sets. So, if everybody's using common data sets that may have these inherent biases in them, we're kind of replicating large-scale biases. This tension between part one and part two and this bias question are very important ones to think through. The good news, though, is that in the last couple years, there's been a growing recognition of the issues we just described. And I think there are now many places that are putting real research effort into these questions about how you think about bias.

David Schwartz: What are best practices for AI, given what we've discussed today about the wide range of applications, the wide range of limitations, and the wide range of challenges before us?

Michael Chui: It is early, so to talk about best practices might be a little bit preliminary. I'll steal a phrase that I once heard from Gary Hamel: we might be talking about next practices, in a certain sense. That said, there a few things that we've observed from leaders who are pioneers and vanguards.

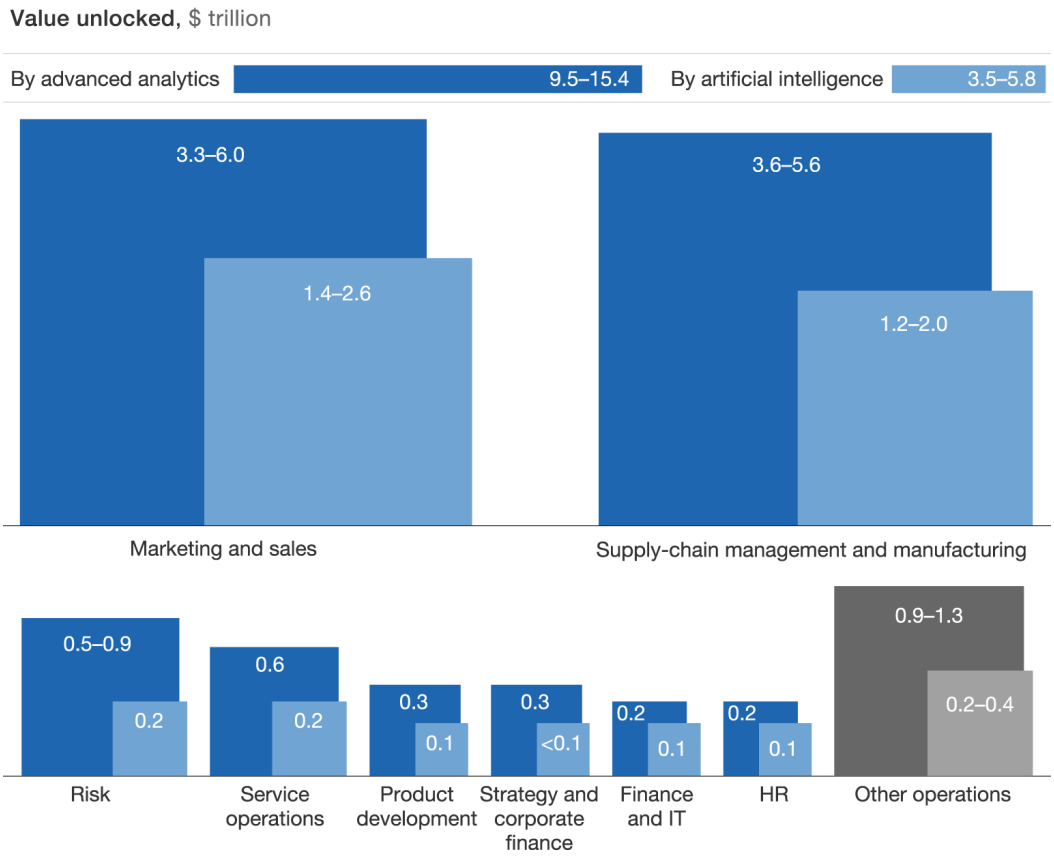
The first thing is one we've described as "get calibrated," but it's really just to start to understand the technology and what's possible. For some of the things that we've talked about today, business leaders over the past few years have had to understand technology more. This is really on the tip of the spear, on the cutting edge. So, really try to understand what's possible in the technology.

Then, try to understand what the potential implications are across your entire business. As we said, these technologies are widely applicable. So, understand where in your business you're deriving value and how these technologies can help you derive value, whether it's marketing and sales, whether it's supply chain, whether it's manufacturing, whether it's in human capital or risk [Exhibit 2].

And then, don't be afraid to be bold. At least experiment. This is a type of technology where it's a learning curve, and the earlier you to start to learn, the faster you'll go up the curve and the

Exhibit 2

Artificial intelligence's impact is likely to be most substantial in marketing and sales as well as supply-chain management and manufacturing, based on our use cases.



Note: Figures may not sum to 100%, because of rounding.

McKinsey&Company | **Source:** McKinsey Global Institute analysis

quicker you'll learn where you can add value, where you can find data, and how you can have a data strategy in order to unlock the data you need to do machine learning. Getting started early—there's really no substitute for that.

James Manyika: The only other thing I would add is something you've been working a lot on, Michael. One of the things that leaders are going to have to understand, or make sure that their teams understand, is this question of which techniques map to which kinds of problems, and also which techniques lead to what kind of value.

We know that the vast majority of the techniques, in the end, are largely classifiers. Knowing that is helpful. Then knowing if the kind of problem sets in your business system are ones that look like classification problems; if so, you have an enormous opportunity. This leads to where you then think about where economic value is and if you have the data available.

There's a much more granular understanding that leaders are going to have to have, unfortunately. The reason why this matters, back to Michael's next-practice point, is that we are already seeing, if you like, a differentiation between those leaders and company who are at the frontier of understanding this and applying these techniques, versus others who are, quite frankly, dabbling—or, at least, paying lip service.

It's worth occasionally as a leader, I would think, visiting or spending time with researchers at the frontier, or at least talking to them, just to understand what's going on and what's not possible. Because this field is moving so quickly. Things that may have been seen as limitations two years ago may not be anymore. And if you're still relying on a conversation you had with an AI scientist two years ago, you may be behind already.

David Schwartz: James and Michael, absolutely fascinating. Thank you for joining us.

James Manyika: Thank you.

Michael Chui: Thank you. □

Michael Chui is a partner of the McKinsey Global Institute (MGI) and is based in McKinsey's San Francisco office, where J**ames Manyika,** chairman and a director of MGI, is a senior partner. **David Schwartz** is a senior editor with McKinsey Publishing and is based in the Stamford office.