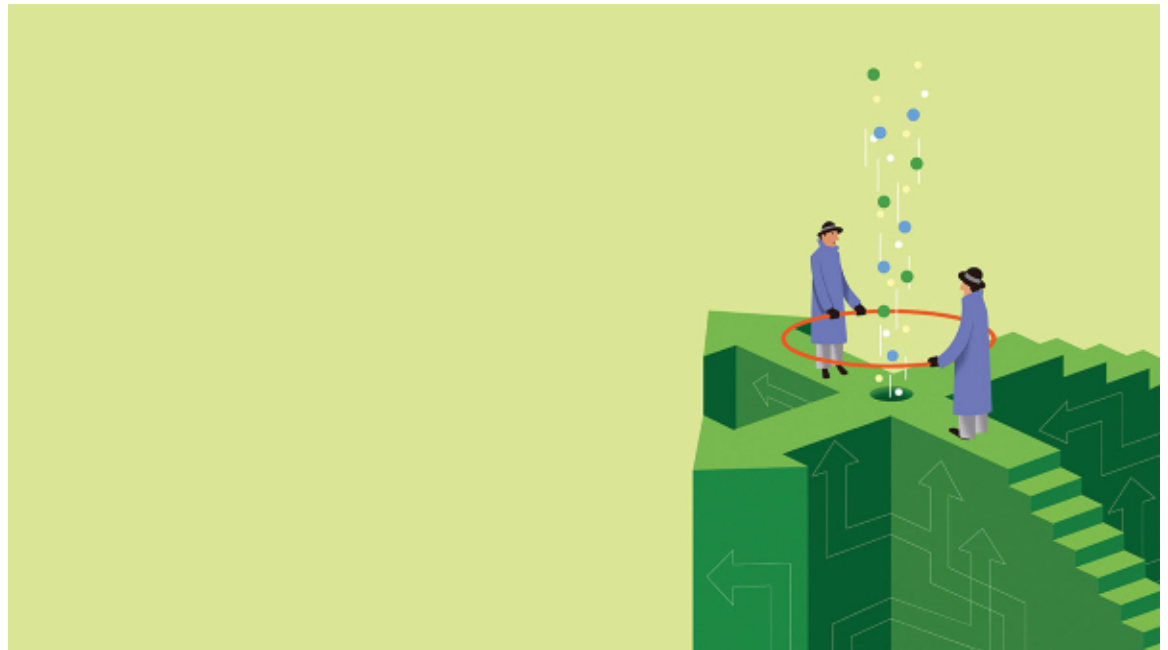


DECISION MAKING

# How AI Will Change the Way We Make Decisions

by Ajay Agrawal, Joshua Gans and Avi Goldfarb  
JULY 26, 2017



With the recent explosion in AI, there has been the understandable concern about its potential impact on human work. Plenty of people have tried to predict which industries and jobs will be most affected, and which skills will be most in demand. (Should you learn to code? Or will AI replace coders too?)

Rather than trying to predict specifics, we suggest an alternative approach. Economic theory suggests that AI will substantially raise the value of human judgment. People who display good judgment will become more valuable, not less. But to understand what good judgment entails and why it will become more valuable, we have to be precise about what we mean.

## **What AI does and why it's useful**

Recent advances in AI are best thought of as a [drop in the cost of prediction](#). By prediction, we don't just mean the future—prediction is about using data that you have to generate data that you don't have, often by translating large amounts of data into small, manageable amounts. For example, using images divided into parts to detect whether or not the image contains a human face is a classic prediction problem. Economic theory tells us that as the cost of machine prediction falls, machines will do more and more prediction.

Prediction is useful because it helps improve decisions. But it isn't the only input into decision-making; the other key input is judgment. Consider the example of a credit card network deciding whether or not to approve each attempted transaction. They want to allow legitimate transactions and decline fraud. They use AI to predict whether each attempted transaction is fraudulent. If such predictions were perfect, the network's decision process is easy. Decline if and only if fraud exists.

However, even the best AIs make mistakes, and that is unlikely to change anytime soon. The people who have run the credit card networks know from experience that there is a trade-off between detecting every case of fraud and inconveniencing the user. (Have you ever had a card declined when you tried to use it while traveling?) And since convenience is the whole credit card business, that trade-off is not something to ignore.

This means that to decide whether to approve a transaction, the credit card network has to know the cost of mistakes. How bad would it be to decline a legitimate transaction? How bad would it be to allow a fraudulent transaction?

Someone at the credit card association needs to assess how the entire organization is affected when a legitimate transaction is denied. They need to trade that off against the effects of allowing a transaction that is fraudulent. And that trade-off may be different for high net worth individuals than for casual card users. No AI can make that call. Humans need to do so. This decision is what we call judgment.

## **What judgment entails**

Judgment is the process of determining what the reward to a particular action is in a particular environment. Judgment is how we work out the benefits and costs of different decisions in different situations.

Credit card fraud is an easy decision to explain in this regard. Judgment involves determining how much money is lost in a fraudulent transaction, how unhappy a legitimate customer will be when a transaction is declined, as well as the reward for doing the right thing and allowing good transactions and declining bad ones. In many other situations, the trade-offs are more complex, and the payoffs are not straightforward. Humans learn the payoffs to different outcomes by experience, making choices and observing their mistakes.

Getting the payoffs right is hard. It requires an understanding of what your organization cares about most, what it benefits from, and what could go wrong.

In many cases, especially in the near term, humans will be required to exercise this sort of judgment. They'll specialize in weighing the costs and benefits of different decisions, and then that judgment will be combined with machine-generated predictions to make decisions.

But couldn't AI calculate costs and benefits itself? In the credit card example, couldn't AI use customer data to consider the trade-off and optimize for profit? Yes, but someone would have had to program the AI as to what the appropriate profit measure is. This highlights a particular form of human judgment that we believe will become both more common and more valuable.

### **Setting the right rewards**

Like people, AIs can also learn from experience. One important technique in AI is reinforcement learning whereby a computer is trained to take actions that maximize a certain reward function. For instance, DeepMind's AlphaGo was trained this way to maximize its chances of winning the game of Go. Games are often easy to apply this method of learning because the reward can be easily described and programmed - shutting out a human from the loop.

But games can be cheated. [As \*Wired\* reports](#), when AI researchers trained an AI to play the boat racing game, CoastRunners, the AI figured out how to maximize its score by going around in circles rather than completing the course as was intended. One might consider this ingenuity of a type, but when it comes to applications beyond games this sort of ingenuity can lead to perverse outcomes.

The key point from the CoastRunners example is that in most applications, the goal given to the AI differs from the true and difficult-to-measure objective of the organization. As long as that is the case, humans will play a central role in judgment, and therefore in organizational decision-making.

In fact, even if an organization is enabling AI to make certain decisions, getting the payoffs right for the organization as a whole requires an understanding of how the machines make those decisions. What types of prediction mistakes are likely? How might a machine learn the wrong message?

Enter Reward Function Engineering. As AIs serve up better and cheaper predictions, there is a need to think clearly and work out how to best use those predictions. Reward Function Engineering is the job of determining the rewards to various actions, given the predictions made by the AI. Being great at it requires having an understanding of the needs of the organization and the capabilities of the machine. (And it is *not* the same as putting a human in the loop to help train the AI.)

Sometimes Reward Function Engineering involves programming the rewards in advance of the predictions so that actions can be automated. Self-driving vehicles are an example of such hard-coded rewards. Once the prediction is made, the action is instant. But as the CoastRunners example illustrates, getting the reward right isn't trivial. Reward Function Engineering has to consider the

possibility that the AI will over-optimize on one metric of success, and in doing so act in a way that's inconsistent with the organization's broader goals.

At other times, such hard-coding of the rewards is too difficult. There may so be many possible predictions that it is too costly for anyone to judge all the possible payoffs in advance. Instead, some human needs to wait for the prediction to arrive, and then assess the payoff. This is closer to how most decision-making works today, whether or not it includes machine-generated predictions. Most of us already do some Reward Function Engineering, but for humans — not machines. Parents teach their children values. Mentors teach new workers how the system operates. Managers give objectives to their staff, and then tweak them to get better performance. Every day, we make decisions and judge the rewards. But when we do this for humans, prediction and judgment are grouped together, and the distinct role of Reward Function Engineering has not needed to be explicitly separate.

As machines get better at prediction, the distinct value of Reward Function Engineering will increase as the application of human judgment becomes central.

Overall, will machine prediction decrease or increase the amount of work available for humans in decision-making? It is too early to tell. On the one hand, machine prediction will substitute for human prediction in decision-making. On the other hand, machine prediction is a complement to human judgment. And cheaper prediction will generate more demand for decision-making, so there will be more opportunities to exercise human judgment. So, although it is too early to speculate on the overall impact on jobs, there is little doubt that we will soon be witness to a great flourishing of demand for human judgment in the form of Reward Function Engineering.

---

**Ajay Agrawal** is the Peter Munk Professor of Entrepreneurship at the University of Toronto's Rotman School of Management and Research Associate at the National Bureau of Economic Research in Cambridge, MA. He is founder of the Creative Destruction Lab, co-founder of The Next AI, and co-founder of Kindred.

---

**Joshua Gans** is professor of strategic management at the Rotman School of Management. His latest book, [The Disruption Dilemma](#), is published by MIT Press.

---

**Avi Goldfarb** is the Ellison Professor of Marketing at the Rotman School of Management, University of Toronto. He is also a Research Associate at the National Bureau of Economic Research, Chief Data Scientist at the Creative Destruction Lab, and Senior Editor at Marketing Science.

---

Copyright 2017 Harvard Business Publishing. All Rights Reserved. Additional restrictions may apply including the use of this content as assigned course material. Please consult your institution's librarian about any restrictions that might apply under the license with your institution. For more information and teaching resources from Harvard Business Publishing including Harvard Business School Cases, eLearning products, and business simulations please visit [hbsp.harvard.edu](http://hbsp.harvard.edu).