NOVEMBER 2017

RISK

Controlling machine-learning algorithms and their biases

Myths aside, artificial intelligence is as prone to bias as the human kind. The good news is that the biases in algorithms can also be diagnosed and treated.

Tobias Baer and Vishnu Kamalnath

Companies are moving quickly to apply machine learning to business decision making. New programs are constantly being launched, setting complex algorithms to work on large, frequently refreshed data sets. The speed at which this is taking place attests to the attractiveness of the technology, but the lack of experience creates real risks. Algorithmic bias is one of the biggest risks because it compromises the very purpose of machine learning. This often-overlooked defect can trigger costly errors and, left unchecked, can pull projects and organizations in entirely wrong directions. Effective efforts to confront this problem at the outset will repay handsomely, allowing the true potential of machine learning to be realized most efficiently.

Machine learning has been in scientific use for more than half a century as a term describing programmable pattern recognition. The concept is even older, having been expressed by pioneering mathematicians in the early 19th century. It has come into its own in the past two decades, with the advent of powerful computers, the Internet, and mass-scale digitization of information. In the domain of artificial intelligence, machine learning increasingly refers to computer-aided decision making based on statistical algorithms generating data-driven insights (see sidebar, "Machine learning: The principal approach to realizing the promise of artificial intelligence").

Among its most visible uses is in predictive modeling. This has wide and familiar business applications, from automated customer recommendations to credit-approval processes. Machine learning magnifies the power of predictive models through great computational force. To create a functioning statistical algorithm by means of a logistic

Machine learning: The principal approach to realizing the promise of artificial intelligence

Artificial intelligence is the science and engineering of automated problem solving. The object is to generate solutions by using computers to mimic the cognitive functions associated with deliberative thought, including perception, reasoning, and learning.

Machine learning is the most prevalent means by which the potential of artificial intelligence is being exploited. The term refers to the ability of computers to detect patterns in large data sets through the application of algorithms. In addition to uncovering potentially powerful insights in the data, computers can be programmed to train themselves to make data-driven predictions.

Predictive modeling, also called supervised learning, is a machine-learning approach that builds pattern-recognition models using sample data with known attributes and outcomes (labeled "training data"). Working from the known patterns, the model can predict outcomes for new observations. The form of data used to predict outcomes can be structured or unstructured, whether or not supervised learning is applied. However, unstructured data can be processed directly only through machine learning; when more traditional techniques such as regression are used, the data scientist must first aggregate unstructured data into structured data based on business rules or independent analyses and procedures.

Deep learning is the most advanced technique for predictive modeling. It connects software-based calculators to form a complex artificial "neural network," often 50 or more layers deep. The simplest predictive-modeling techniques are regression modeling and simple decision trees. More advanced techniques include random forests (a more complex and sensitive decision-tree model) and support vector machines (for sophisticated data classification).

regression, for example, missing variables must be replaced by assumed numeric values (a process called imputation). Machine-learning algorithms are often constructed to interpret "missing" as a possible value and then proceed to develop the best prediction for cases where the value is missing. Machine learning is able to manage vast amounts of data and detect many more complex patterns within them, often attaining superior predictive power.

In credit scoring, for example, customers with a long history of maintaining loans without delinquency are generally determined to be of low risk. But what if the mortgages these customers have been maintaining were for years supported by substantial tax benefits that are set to expire? A spike in defaults may be in the offing, unaccounted for in the statistical risk model of the lending institution. With access to the right data and guidance by subject-matter experts, predictive machine-learning models could find the hidden patterns in the data and correct for such spikes.

The persistence of bias

In automated business processes, machine-learning algorithms make decisions faster than human decision makers and at a fraction of the cost. Machine learning also promises to improve decision quality, due to the purported absence of human biases. Human decision makers might, for example, be prone to giving extra weight to their personal experiences. This is a form of bias known

as anchoring, one of many that can affect business decisions. Availability bias is another. This is a mental shortcut (heuristic) by which people make familiar assumptions when faced with decisions. The assumptions will have served adequately in the past but could be unmerited in new situations. Confirmation bias is the tendency to select evidence that supports preconceived beliefs, while loss-aversion bias imposes undue conservatism on decision-making processes.

Machine learning is being used in many decisions with business implications, such as loan approvals in banking, and with personal implications, such as diagnostic decisions in hospital emergency rooms. The benefits of removing harmful biases from such decisions are obvious and highly desirable, whether they come in financial, medical, or some other form.

Some machine learning is designed to emulate the mechanics of the human brain, such as deep learning, with its artificial neural networks. If biases affect human intelligence, then what about artificial intelligence? Are the machines biased? The answer, of course, is yes, for some basic reasons. First, machine-learning algorithms are prone to incorporating the biases of their human creators. Algorithms can formalize biased parameters created by sales forces or loan officers, for example. Where machine learning predicts behavioral outcomes, the necessary reliance on historical criteria will reinforce past biases, including stability bias. This is the tendency to discount the possibility of significant change—for example, through substitution effects created by innovation. The severity of this bias can be magnified by machine-learning algorithms that must assume things will more or less continue as before in order to operate. Another basic bias-generating factor is incomplete data. Every machine-learning algorithm operates wholly within the world defined by the data that were used to calibrate it. Limitations in the data set will bias outcomes, sometimes severely.

Predicting behavior: 'Winner takes all'

Machine learning can perpetuate and even amplify behavioral biases. By design, a social-media site filtering news based on user preferences reinforces natural confirmation bias in readers. The site may even be systematically preventing perspectives from being challenged with contradictory evidence. The self-fulfilling prophecy is a related by-product of algorithms. Financially sound companies can run afoul of banks' scoring algorithms and find themselves without access to working capital. If they are unable to sway credit officers with factual logic, a liquidity crunch could wipe out an entire class of businesses. These examples reveal a certain "winner takes all" outcome that affects those machine-learning algorithms designed to replicate human decision making.

Data limitations

Machine learning can reveal valuable insights in complex data sets, but data anomalies and errors can lead algorithms astray. Just as a traumatic childhood accident can cause lasting behavioral distortion in adults, so can unrepresentative events cause machine-learning algorithms to go off course. Should a series of extraordinary weather events or fraudulent actions trigger spikes in default rates, for example, credit scorecards could brand a region as "high risk" despite the absence of a permanent structural cause. In such cases, inadequate algorithms will perpetuate bias unless corrective action is taken.

Companies seeking to overcome biases with statistical decision-making processes may find that the data scientists supervising their machine-learning algorithms are subject to these same biases. Stability biases, for example, may cause data scientists to prefer the same data that human decision makers have been using to predict outcomes. Cost and time pressures, meanwhile, could deter them from collecting other types of data that harbor the true drivers of the outcomes to be predicted.

The problem of stability bias

Stability bias—the tendency toward inertia in an uncertain environment—is actually a significant problem for machine-learning algorithms. Predictive models operate on patterns detected in historical data. If the same patterns cease to exist, then the model would be akin to an old railroad timetable—valuable for historians but not useful for traveling in the here and now. It is frustratingly difficult to shape machine-learning algorithms to recognize a pattern that is not present in the data, even one that human analysts know is likely to manifest at some point. To bridge the gap between available evidence and self-evident reality, synthetic data points can be created. Since machine-learning algorithms try to capture patterns at a very detailed level, however, every attribute of each synthetic data point would have to be crafted with utmost care.

In 2007, an economist with an inkling that credit-card defaults and home prices were linked would have been unable to build a predictive model showing this relationship, since it had not yet appeared in the data. The relationship was revealed, precipitously, only when the financial crisis hit and housing prices began to fall. If certain data limitations are permitted to govern modeling choices, seriously flawed algorithms can result. Models will be unable to recognize obviously real but unexpected changes. Some US mortgage models designed before the financial crisis could not mathematically accept negative changes in home prices. Until negative interest rates appeared in the real world, they were statistically unrecognized and no machine-learning algorithm in the world could have predicted their appearance.

Addressing bias in machine-learning algorithms

As described in a previous article in *McKinsey on Risk*,[1] companies can take measures to eliminate bias or protect against its damaging effects in human decision making. Similar countermeasures can protect against algorithmic bias. Three filters are of prime importance.

First, users of machine-learning algorithms need to understand an algorithm's shortcomings and refrain from asking questions whose answers will be invalidated by algorithmic bias. Using a machine-learning model is more like driving a car than riding an elevator. To get from point A to point B, users cannot simply push a button; they must first learn operating procedures, rules of the road, and safety practices.

Second, data scientists developing the algorithms must shape data samples in such a way that biases are minimized. This step is a vital and complex part of the process and worthy of much deeper consideration than can be provided in this short article. For the moment, let us remark that available historical data are often inadequate for this purpose, and fresh, unbiased data must be generated through a controlled experiment.

Finally, executives should know when to use and when not to use machine-learning algorithms. They must understand the true values involved in the trade-off: algorithms offer speed and convenience, while manually crafted models, such as decision trees or logistic regression—or for that matter even human decision making—are approaches that have more flexibility and transparency.

What's in your black box?

From a user's standpoint, machine-learning algorithms are black boxes. They offer quick and easy solutions to those who know little or nothing of their inner workings. They should be applied with discretion, but knowing enough to exercise discretion takes effort. Business users seeking to avoid harmful applications of algorithms are a little like consumers seeking to eat healthy food. Health-conscious consumers must study literature on nutrition and read labels in order to avoid excess calories, harmful additives, or dangerous allergens. Executives and practitioners will likewise have to study the algorithms at the core of their business and the problems they are designed to resolve.

They will then be able to understand monitoring reports on the algorithms, ask the right questions, and challenge assumptions.

In credit scoring, for example, built-in stability bias prevents machine-learning algorithms from accounting for certain rapid behavioral shifts in applicants. These can occur if applicants recognize the patterns that are being punished by models. Salespeople have been known to observe the decision patterns embedded in algorithms and then coach applicants by reverse-engineering the behaviors that will maximize the odds of approval.

A subject that frequently arises as a predictor of risk in this context is loan tenor. Riskier customers generally prefer longer loan tenors, in recognition of potential difficulties in repayment. Many low-risk customers, by contrast, aim to minimize interest expense by choosing shorter tenors. A machine-learning algorithm would jump on such a pattern, penalizing applications for longer tenors with a higher risk estimate. Soon salespeople would nudge risky applicants into the approval range of the credit score by advising them to choose the shortest possible tenor. Burdened by an exceptionally high monthly installment (due to the short tenor), many of these applicants will ultimately default, causing a spike in credit losses.

Astute observers can thus extract from the black box the variables with the greatest influence on an algorithm's predictions. Business users should recognize that in this case loan tenor was an influential predictor. They can either remove the variable from the algorithm or put in place a safeguard to prevent a behavioral shift. Should business users fail to recognize these shifts, banks might be able to identify them indirectly, by monitoring the distribution of monthly applications by loan tenor. The challenge here is to establish whether a marked shift is due to a deliberate change in behavior by applicants or to other factors, such as changes in economic conditions or a bank's promotional strategy. In one way or the other, sound business judgment therefore is indispensable.

Squeezing bias out of the development sample

Tests can ensure that unwanted biases of past human decision makers, such as gender biases, for example, have not been inadvertently baked into machine-learning algorithms. Here a challenge lies in adjusting the data such that the biases disappear.

One of the most dangerous myths about machine learning is that it needs no ongoing human intervention. Business users would do better to view the application of machine-learning algorithms like the creation and tending of a garden. Much human oversight is needed. Experts with deep machine-learning knowledge and good business judgment are like experienced gardeners, carefully nurturing the plants to encourage their organic growth. The data scientist knows that in machine learning the answers can be useful only if we ask the right questions.

In countering harmful biases, data scientists seek to strengthen machine-learning algorithms where it most matters. Training a machine-learning algorithm is a bit like building muscle mass. Fitness trainers take great pains in teaching their clients the proper form of each exercise so that only targeted muscles are worked. If the hips are engaged in a motion designed to build up biceps, for example, the effectiveness of the exercise will be much reduced. By using stratified sampling and optimized observation weights, data scientists ensure that the algorithm is most powerful for those decisions in which the business impact of a prediction error is the greatest. This cannot be done automatically, even by advanced machine-learning algorithms such as boosting (an algorithm designed to reduce algorithmic bias). Advanced algorithms can correct for a statistically defined concept of error, but they cannot distinguish errors with high business impact from those of negligible importance. Another example of the many statistical techniques data scientists can deploy to protect algorithms

from biases is the careful analysis of missing values. By determining whether the values are missing systematically, data scientists are introducing "hindsight bias." This use of bias to fight bias allows the algorithm to peek beyond its data-determined limitations to the correct answer. The data scientists can then decide whether and how to address the missing values or whether the sample structure needs to be adjusted.

Deciding when to use machine-learning algorithms

An organization considering using an algorithm on a business problem should be making an explicit choice based on the cost-benefit trade-off. A machine-learning algorithm will be fast and convenient, but more familiar, traditional decision-making processes will be easier to build for a particular purpose and will also be more transparent. Traditional approaches include human decision making or handcrafted models such as decision trees or logistic-regression models—the analytic workhorses used for decades in business and the public sector to assign probabilities to outcomes. The best data scientists can even use machine-learning algorithms to enhance the power of hand-crafted models. They have been able to build advanced logistic-regression models with predictive power approaching that of a machine-learning algorithm.

Three questions can be considered when deciding to use machine-learning algorithms:

- *How soon do we need the solution?* The time factor is often of prime importance in solving business problems. The optimal statistical model may be obsolete by the time it is completed. When the business environment is changing fast, a machine-learning algorithm developed overnight could far outperform a superior traditional model that is months in the making. For this reason, machine-learning algorithms are preferred for combating fraud. Defrauders typically act quickly to circumvent the latest detection mechanisms they encounter. To defeat fraud, organizations need to deploy algorithms that adjust instantaneously, the moment the defrauders change their tactics.

- *What insights do we have?* The superiority of the handcrafted model depends on the business insights embedded in it by the data scientist. If an organization possesses no insights, then the problem solving will have to be guided by the data. At this point, a machine-learning algorithm might be preferred for its speed and convenience. However, rather than blindly trusting an algorithm, an organization in this situation could decide that it is better to bring in a consultant to help develop value-adding business insights.

- *Which problems are worth solving?* One of the promises of machine learning is that it can address problems that were once unrecognized or thought to be too costly to solve with a handcrafted model. Decision making on these problems has been heretofore random or unconscious. When reconsidering such problems, organizations should identify those with significant bottom-line business impact and then assign their best data scientists to work on them.

In addition to these considerations, companies implementing large-scale machine-learning programs should make appropriate organizational and cultural changes to support them. Everyone within the scope of the programs should understand and trust the machine-learning models—only then will maximum impact be achieved.

Implementation: Standards, validation, knowledge

How would a business go about implementing these recommendations? The practical application and debiasing of machine-learning algorithms should be governed by a conscious and eventually systematic process throughout the organization. While not as stringent and formal, the approach is related to

mature model development and validation processes by which large institutions are gaining strategic control of model proliferation and risk. Three building blocks are critically important for implementation:

- *Business-based standards for machine-learning approvals.* A template should be developed for model documentation, standardizing the process for the intake of modeling requests. It should include the business context and prompt requesters with specific questions on business impact, data, and cost-benefit trade-offs. The process should require active user participation in the drive to find the most suitable solution to the business problem (note that passive checklists or guidelines, by comparison, tend to be ignored). The model's key parameters should be defined, including a standard set of analyses to be run on the raw data inputs, the processed sample, and the modeling outputs. The model should be challenged in a discussion with business users.

- *Professional validation of machine-learning algorithms.* An explicit process is needed for validating and approving machine-learning algorithms. Depending on the industry and business context—especially the economic implication of errors—it may not have to be as stringent as the formal validation of banks' risk models by internal validation teams and regulators. However, the process should establish validation standards and an ongoing monitoring program for the new model. The standards should account for the characteristics of machine-learning models, such as automatic updates of the algorithm whenever fresh data are captured. This is an area where most banks still need to develop appropriate validation and monitoring standards. If algorithms are updated weekly, for example, validation routines must be completed in hours and days rather than weeks and months. Yet it is also extremely important to put in place controls that alert users to potential sudden or creeping bias in fresh data.

- *A culture for continuous knowledge development.* Institutions should invest in developing and disseminating knowledge on data science and business applications. Machine-learning applications should be continuously monitored for new insights and best practices, in order to create a culture of knowledge enhancement and to keep people informed of both the difficulties and successes that come with using such applications.

Creating a conscious, standards-based system for developing machine-learning algorithms will involve leaders in many judgment-based decisions. For this reason, debiasing techniques should be deployed to maximize outcomes. An effective technique in this context is a "premortem" exercise designed to pinpoint the limitations of a proposed model and help executives judge the business risks involved in a new algorithm.

Sometimes lost in the hype surrounding machine learning is the fact that artificial intelligence is as prone to bias as the real thing it emulates. The good news is that biases can be understood and managed—if we are honest about them. We cannot afford to believe in the myth of machine-perfected intelligence. Very real limitations to machine learning must be constantly addressed by humans. For businesses, this means the creation of incremental, insights-based value with the aid of well-monitored machines. That is a realistic algorithm for achieving machine-learning impact.

[1] Tobias Baer, Sven Heiligtag, and Hamid Samandari, "The business logic in debiasing," May 2017, McKinsey.com.

Tobias Baer is a partner in McKinsey's Taipei office, and **Vishnu Kamalnath** is an analytics specialist in the North American Knowledge Center in Waltham, Massachusetts.