THE RISK OF MACHINE LEARNING BIAS (AND HOW TO PREVENT IT)

As promising as machine-learning technology is, it can also be susceptible to unintended biases that require careful planning to avoid

Chris DeBrusk

MANY COMPANIES ARE turning to machine learning to review vast amounts of data, from evaluating credit for loan applications, to scanning legal contracts for errors, to looking through employee communications with customers to identify bad conduct. New tools allow developers to build and deploy machine-learning engines more easily than ever: Amazon Web Services recently launched a “machine learning in a box” offering called SageMaker that non-engineers can leverage to build sophisticated machine-learning models, and Microsoft Azure’s machine-learning platform, Machine Learning Studio, requires no coding skills.

But while machine-learning algorithms enable companies to realize new efficiencies, they are as susceptible as any system to the “garbage in, garbage out” syndrome. In the case of self-learning systems, the type of “garbage” is biased data. Left unchecked, feeding biased data to self-learning systems can lead to unintended and sometimes dangerous outcomes.

In 2016, for example, an attempt by Microsoft to converse with millennials using a chat bot plugged into Twitter famously created a racist machine that switched from tweeting that “humans are super cool” to praising Hitler and spewing out misogynistic remarks. This scary conclusion to a one-day experiment resulted from a very straightforward rule about machine learning – the models learn exactly what they are taught. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a machine-learning system that makes recommendations for criminal sentencing, is also proving imperfect at predicting which people are likely to reoffend because it was trained on incomplete data. Its training model includes race as an input parameter, but not more extensive data points such as past arrests. As a result, it has an inherent racial bias that is difficult to accept as either valid or just.

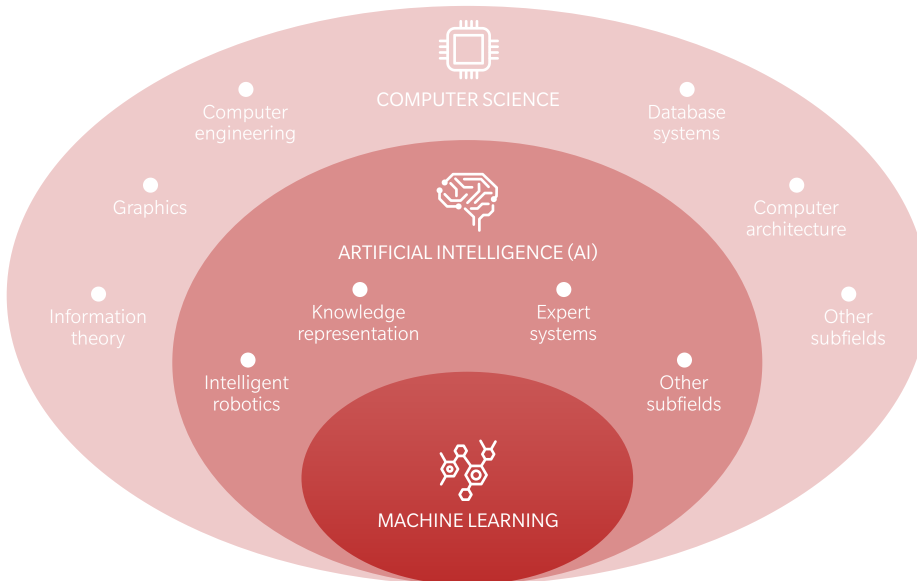
These are just two of many cases of machine-learning bias. Yet there are many more potential ways in which machines can be taught to do something immoral, unethical, or just plain wrong.

These examples serve to underscore why it is so important for managers to guard against the potential reputational and regulatory risks that can result from biased data, in addition to figuring out how and where machine-learning models should be deployed to begin with. Best practices are emerging that can help to prevent machine-learning bias. Below, we examine a few.

Consider bias when selecting training data. Machine-learning models are, at their core, predictive engines. Large data sets train machine-learning models to predict the future based on the past. Models can read masses of text and understand intent, where intent is known. They can learn to spot differences – between, for instance, a cat and a dog – by consuming millions of pieces of data, such as correctly labeled animal photos.

EXHIBIT 1: A TAXONOMY OF MACHINE-LEARNING TERMS

Navigating artificial intelligence and machine-learning concepts can sometimes be daunting. Below is a breakdown of some of the key terms.



ARTIFICIAL INTELLIGENCE (AI)

is a scientific field within Computer Science, focusing on the study of computer systems that can perform tasks and solve problems that require human intelligence

MACHINE LEARNING

is a field within AI that focuses on a particular class of algorithms that can learn from data without being explicitly programmed

There are three main ways a machine can learn from data

Each category of machine learning is effective in tackling particular kinds of tasks and problems

There is a wide range of mathematical techniques that can be used to develop machine-learning models

Supervised machine learning

Mapping inputs to labeled outputs

Commonly used in classification and regression problems such as:

Natural language processing
Image recognition
Financial forecasting

Decision tree | Random forest | Gradient boosting | Neural network | Support vector machine | Regularized regression

Unsupervised machine learning

Finding patterns in unlabeled input data

Commonly used in segmentation and clustering problems such as:

Pattern and trend recognition
Customer segmentation
Transaction monitoring

Clustering | Dimensionality reduction | Gaussian mixture model | Principal component analysis | Independent component analysis

Reinforcement learning

Performing actions to maximize rewards

Similar to supervised learning, but reward mechanism in place instead of labelled output:

Board and video games
Robotics

Q-learning | Temporal difference algorithm | Neural network | State-action-reward-state-action

Note: Not comprehensive, key elements shown
Source: Oliver Wyman analysis

The advantage of machine-learning models over traditional statistical models is their ability to quickly consume enormous numbers of records and thereby more accurately make predictions. But since machine-learning models predict exactly what they have been trained to predict, their forecasts are only as good as the data used for their training.

For example, a machine-learning model designed to predict the risk of business loan defaults may advise against extending credit to companies with strong cash flows and solid management teams if it draws a faulty connection – based on data from loan officers' past decisions – about loan defaults by businesses run by people of a certain race or in a particular zip code. A machine-learning model used to scan reams of résumés or applications to schools might mistakenly screen out female applicants if the historical data used to train it reflects past decisions that resulted in few women being hired or admitted to a college.

These types of biases are especially pervasive in data sets based on decisions made by a relatively small number of people. As a best practice, managers must always keep in mind that if humans are involved in decisions, bias always exists – and the smaller the group, the greater the chance that the bias is not overridden by others.

Root out bias. To address potential machine-learning bias, the first step is to honestly and openly question what preconceptions could currently exist in an organization's processes, and actively hunt for how those biases might manifest themselves in data. Since this can be a delicate issue, many organizations bring in outside experts to challenge their past and current practices.

Once potential biases are identified, companies can block them by eliminating problematic data or removing specific components of the input data set. Managers for a credit card company, for example, when considering how to address late payments or defaults, might initially build a model with data such as zip codes, type of car driven, or certain first names – without acknowledging that these data points can correlate with race or gender. But that data should be stripped, keeping only data directly relevant to whether or not customers will pay their bills, such as data on credit scores or employment and salary information. That way, companies can build a solid machine-learning model to predict likelihood of payment and determine which credit card customers should be offered more flexible payment plans and which should be referred to collection agencies.

A company can also expand the training data set with more information to counterweight potentially problematic data. Some companies, for example, have started to include social media data when evaluating the risk of a customer or client committing a financial crime. A machine-learning algorithm may flag a customer as high risk if he or she starts to post photos on social media from countries with potential terrorist or money-laundering connections. This conclusion can be tested and overridden, though, if a user's nationality, profession, or travel proclivities are included to allow for a native visiting their home country or a journalist or businessperson on a work trip.

Regardless of which approach is used, as a best practice, managers must not take data sets at face value. It is safe to assume that bias exists in all data. The question is how to identify it and remove it from the model.

It is safe to assume that bias exists in all data. The question is how to identify it and remove it from the model.

Counter bias in "dynamic" data sets. Another challenge for machine-learning models is to avoid bias where the data set is dynamic. Since machine-learning models are trained on events that have already happened, they cannot predict outcomes based on behavior that has not been statistically measured. For example, even though machine learning is extensively used in fraud detection, fraudsters can outmaneuver models by devising new ways to steal or escape detection. Employees can hide bad behavior from machine-learning tools used to identify bad conduct by using underhanded techniques like conversing in code.

To attempt to draw new conclusions from current information, some companies use more experimental, cognitive, or artificial intelligence techniques that model potential scenarios. For example, to outsmart money launderers, banks may conduct so-called war games with ex-prosecutors and investigators to discover how they would beat their system. That data is then used to handcraft a more up-to-date machine-learning algorithm.

But even in this situation, managers risk infusing bias into a model when they introduce new parameters. For example, social media data, such as pictures posted on Facebook and Twitter, is increasingly being used to drive predictive models. But a model that ingests this type of data might introduce irrelevant biases into its predictions, such as correlating people wearing blue shirts with improved creditworthiness.

To avoid doing so, managers must ensure that the new parameters are comprehensive and empirically tested – another best practice. Otherwise, those parameters might skew the model, especially in areas where data is poor. Insufficient data could impact, say, credit decisions for classes of borrowers to whom a bank has never lent to previously but plans to in the future.

Balance transparency against performance. One temptation with machine learning is to throw increasingly large amounts of data at a sophisticated training infrastructure and allow the machine to "figure it out." For example, public cloud companies have recently released comprehensive tools that use automated algorithms instead of an expert data scientist to train and determine the parameters intended to optimize machine-learning models.

While this is a powerful method for building complex predictive algorithms quickly and at lower cost, it also comes with the downside of limited visibility and the risk of the "machine running wild" and having an unconscious bias due to training data that is extraneous (like the blue shirt bias described above). The other challenge is that it is very difficult to explain how complex machine-learning models actually work, which is problematic in industries that are heavily regulated.

One of the potential options to address this risk is to take a staged approach to increasing the sophistication of the model and making a conscious decision to progress at every stage.

A good example is a process used by a major bank in building a model that attempted to predict whether a mortgage customer was about to refinance, with the goal of making a direct offer to that customer and ideally retaining their business. The bank started with a simple regression-based model that tested its ability to predict when customers would refinance. It then created a set of more sophisticated "challenger" models that used more advanced machine-learning techniques and were more precise. By confirming that the challenger models were more accurate than the base regression model, bank managers became comfortable that their more complex and opaque machine-learning approach was operating in line with expectations and not propagating unintended biases. The process also enabled them to verify that the machine-learning tool's balance between transparency and sophistication was in line with what is expected in the highly regulated financial services industry.

CAREFUL PLANNING IS A NECESSITY

It is tempting to assume that, once trained, a machine-learning model will continue to perform without oversight. In reality, the environment in which the model is operating is constantly changing, and managers need to periodically retrain models using new data sets.

Machine learning is one of the most exciting technical capabilities with real-world business value to have emerged over the past decade. When combined with big data technology and the massive computing capability available via the public cloud, machine learning promises to change how people interact with technology, and potentially entire industries. But as promising as machine-learning technology is, it requires careful planning to avoid unintended biases.

Creators of the machine-learning models that will drive the future must consider how bias might negatively impact the effectiveness of the decisions the machines make. Otherwise, managers risk undercutting machine learning's potentially positive benefits by building models with a biased "mind of their own."

Chris DeBrusk is a New York-based partner in Oliver Wyman's Financial Services and Digital practices.

This article first appeared in MIT Sloan Management Review on March 26, 2018.