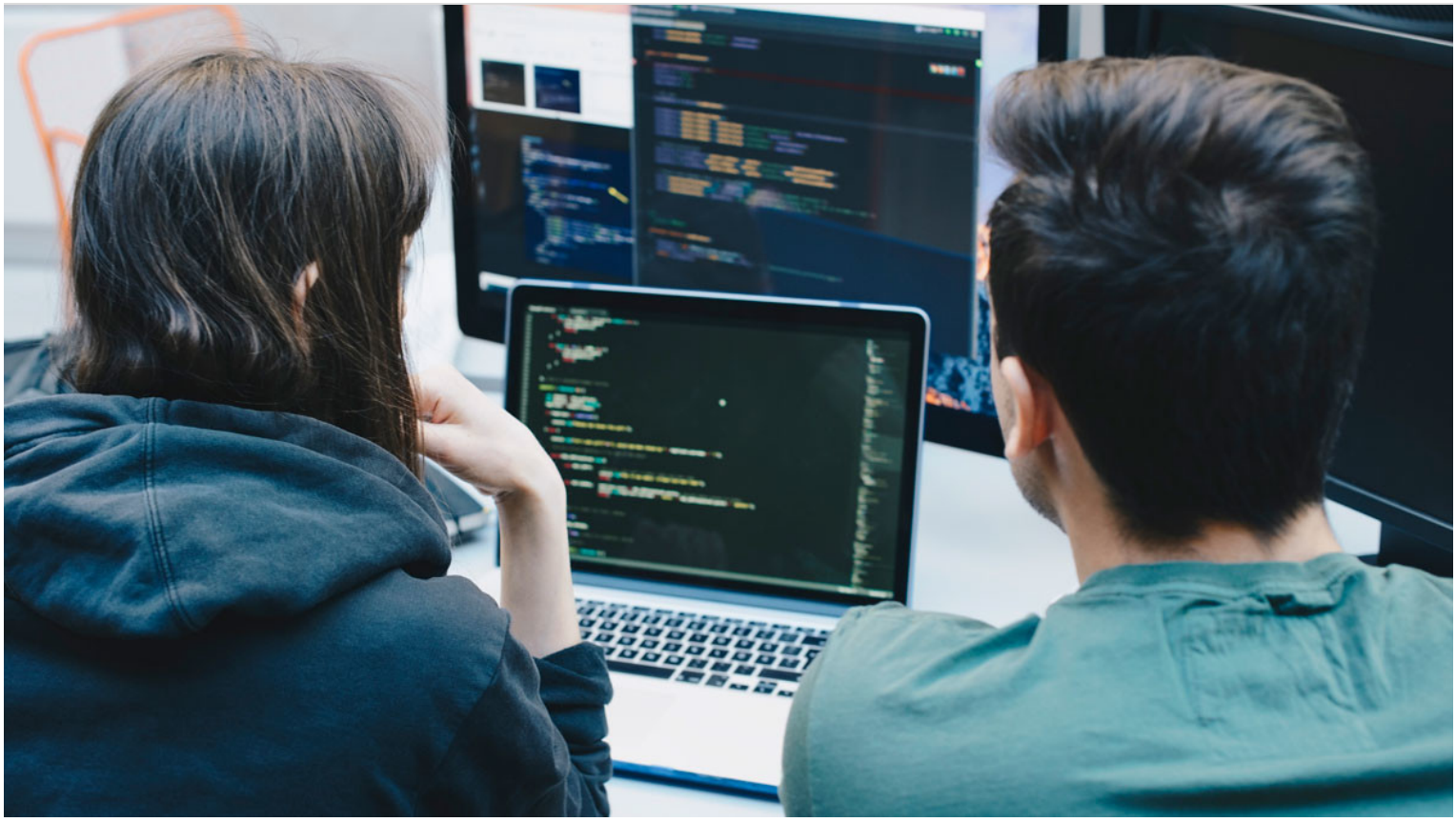
Harvard Business Review

TECHNOLOGY

When Is It Important for an Algorithm to Explain Itself?

by Kathryn Hume

JULY 06, 2018



MASKOT/GETTY IMAGES

Many efforts to apply machine learning get stuck due to concerns about the "black box" — that is, the lack of transparency around why a system does what it does. Sometimes this is because people want to understand why some prediction was made before they take life-altering actions, as when a computer vision system indicates a 95% likelihood of cancer from an x-ray of a patient's lung. Sometimes it's because technical teams need to identify and resolve bugs without disrupting the entire system. And now that the General Data Protection Regulation (GDPR) is in effect, businesses that

handle consumer data are required to explain how automated systems make decisions, especially those that significantly affect individual lives, like allocating credit or hiring a candidate for a job. While GDPR only applies in Europe, businesses around the world anticipate that similar changes are coming and so are revisiting governance efforts.

If you search around the internet, you'll find that most writing about algorithmic explainability falls into two camps. Advocates for rapid technology adoption often argue that humans are no better at explaining decisions than machines, and so we should table the question to accelerate innovation. These rhetorical arguments do not help professionals responsible for regulatory compliance. On the other hand, critics demand stringent requirements for transparency and vilify a "move fast and break things" culture. These arguments can stifle adoption, as not all machine learning use cases require the same level of oversight and accountability — some decisions are more important to be able to explain than others.

To succeed with applied machine learning, you have to step back and break down the problem. What does it mean for a mathematical or statistical procedure to be "appropriate" (as GDPR requires)? Do different use cases require different types of explanations? Who should be involved in decisions regarding business impact, regulatory compliance, technical approach, and even ethical values when companies integrate machine learning into business operations?

Let's start by unpacking why a seemingly straightforward idea like a right to an explanation is hard to understand and implement in practice.

As with any technology, when you start a machine learning project you have to decide whether to build or buy. Working with a vendor complicates transparency because many software companies choose not to disclose what algorithms they use or the data they use to train them. Often, the reason given is to protect intellectual property or prevent a security breach. There's also a complexity issue: If the vendor uses multiple public and private data sets to train their system, think about how difficult it would be to have auditing mechanisms to keep track of exactly what went into making a decision!

If you're not using a vendor, but choosing to build something in-house, you have to decide whether you need only be able to explain what procedures you'll be using — for example, the types of data and types of models — or whether you want to be able to explain the inner workings of a mathematical model.

The language in GDPR implies that it's the procedure that requires explanation. Recital 71 says that "fair and transparent processing" means auditing how data is collected, keeping data accurate, securing data, and taking measures to identify and prevent any discriminatory effects. The focus is on data collection and integrity; statistical models need to be "appropriate." None of these steps are trivial, but they are often overlooked in debates around explainability because there

is so much focus on algorithms and models. For example, bias can creep into an algorithm at many points in the system. Your business may have historically underserved some ethnic population, so you may have collected little data about them. Ethnic and demographic communities may be tightly correlated with location data, leading a seemingly innocuous variable like GPS location to be proxy for ethnic discrimination. Once in production, models often encounter edge cases — situations, data, or individuals that aren't enough like the data they've been trained on. It's important to monitor for bias both before and after a system goes into production, and to take action to address unintended treatment.

One kind of explanation is to clarify the outcomes a system is designed to optimize for. In the example of an online credit application system, holding a system accountable would mean monitoring to ensure that denials were not correlated to protected attributes like ethnic background. The limitations of this outcomes-focused approach is that there is less insight into what an individual would need to do to intervene to change a decision in the future. An intervention-focused approach requires insight into the inner workings of a model. For example: "You didn't qualify because you did not pay your last three rent checks. If you pay the next four in a row, your score will be high enough to pass our threshold score of 75%."

When it's important to understand the logic of a statistical model, we hit different challenges.

As I hinted at in my article about identifying machine learning opportunities, different machine learning algorithms are more and less easy to explain. A linear regression of the form $y = mx + b$ isn't too hard to explain: we only have to track m to know how x (input) relates to y (output). But what if "m" is shorthand for millions of relationships, defining complex functions in architectures? With deep learning we lose the ability to pinpoint how inputs relate to outputs because the number of variables included and the relationships between them become too complex to describe. So, for example, a deep neural network is able to indicate a 95% chance that an individual will default on a loan, but cannot articulate what aspects in the data formed that score. It's a trade-off, as more complex algorithms unlock capabilities simpler statistical models like linear regression cannot handle — but at the cost of explainability. (It's also worth remembering that when data scientists build simpler algorithms that may be easier to explain, they also bring with them biases and assumptions that influence what they see in the data; these subjective biases are hard to identify and control using technology.)

A final challenge in explainability is to make it clear what the model actually optimizes for. An ideal credit card customer is one who will frequently use the card he or she signs up for (long-term outcome), not just the person who accepts the credit card offer (short-term outcome). People who click on display ads aren't often customers with high lifetime value, and most digital marketing efforts are only able to use clickstream data as proxies for direct sales. It's hard to measure and get feedback on long-term outcomes, but these known unknowns can be the most valuable for a system's performance.

This may seem daunting, but if the right people ask the right questions at the right time to inform a series of judgment calls and decisions, things become tractable.

To start, non-technical stakeholders involved in a machine learning project need some training to build intuitions about how statistical systems work. They don't need to code or to be data scientists, but they do need to appreciate that machine learning systems output correlations and not causes. They need to appreciate that a minority group not well represented in a data set may receive unfair treatment from an algorithm, not because of any malice on the part of the data scientists, but because models tend to learn relationships that help predict large parts of the dataset, at the expense of accuracy with respect to less well-represented examples.

Next, during pre-project discussions, a diverse group of stakeholders from the business, data science, IT, privacy, compliance should have a seat at the table. (Companies should also consider explicitly making it someone's role to question the algorithm, like the "red teams" sometimes used in high-stakes decision-making.) It's important to get clear on regulatory requirements or ethical risks before any work begins to avoid sunk costs on interesting applications that won't meet requirements under new regulations like GDPR or risk denigrating consumer trust.

These cross-functional design groups should consider questions like:

What type of accountability matters for the use case? Explainability is not always important. For example, if a law firm uses machine learning to find documents relevant for a case, what matters is that they don't miss something important, not explaining why one document is relevant and another isn't. Here, the right metric for data scientists to focus on is known as "recall," the fraction of relevant instances that have been retrieved over the total amount of relevant instances, across a document set. The data science team should embed this into their model testing and quality assurance processes.

Where does a particular machine learning model sit in the entire business process? A business analyst should map out the end-to-end business process. Often one process actually includes many machine learning models with different explainability requirements. For example, a bank using machine learning to acquire new credit card customers will have at least two models: one to evaluate risk and approve the card (which requires greater explainability) and another to predict propensity to convert and to personalize offers (which requires lower explainability). Compliance functions should inform business analysts of the regulatory requirements at each phase in the business process and data scientists should keep these restrictions in mind as opposed to only selecting the machine learning technique that has the best performance on a task.

What processes will we use to govern outcomes? Machine learning systems are optimization tools, and one way to govern them is to shift from explaining what features in data led to which outcomes to declaring a higher-level policy on desired outcomes and holding systems accountable to achieving that policy. Here, data scientists should have responsibility to evaluate their models for bias towards sensitive data types like gender or ethnic background during quality assurance and, most importantly, after the model goes live. Statistical systems do well in the middle of the bell curve where they have lots of data, but can produce unexpected results on less well-represented cases or new behavior. Someone should be made responsible to audit and monitor model performance over time and identify any actions against business policy. The technical, business, and compliance teams should meet regularly to review performance and adjust the model to achieve fair outcomes. The business should document how frequently models are updated and have a process to communicate this and how it impacts predictions and any changes to consumers impacted by the system.

Much of the conversation around explainability and interpretability focuses narrowly on the inner workings of machine learning models, leading to fear of black boxes or rhetorical arguments that humans are no better at explaining their behavior and decisions than the most opaque machine. For businesses to succeed with machine learning, they have to step back and break down the problem, considering the impact of systems holistically and thinking critically about what meaningful accountability entails for different use cases. In some cases, individuals will indeed require more direct explanations, be that for psychological comfort (being diagnosed with cancer) or to intervene to change an outcome (modifying actions to get a housing loan in the future after one has been denied). But there are many processes that can be governed by setting policies for desired outcomes, monitoring results to track discrepancies, and updating models or data collection procedures to improve future results. Getting clear on what matters and making judgment calls on how much error a business can accept is the skill business leaders need to develop.

Kathryn Hume is vice president of product and strategy at integrate.ai, a Toronto-based startup.

This article is about TECHNOLOGY

⊕ **FOLLOW** THIS TOPIC

Related Topics: DATA