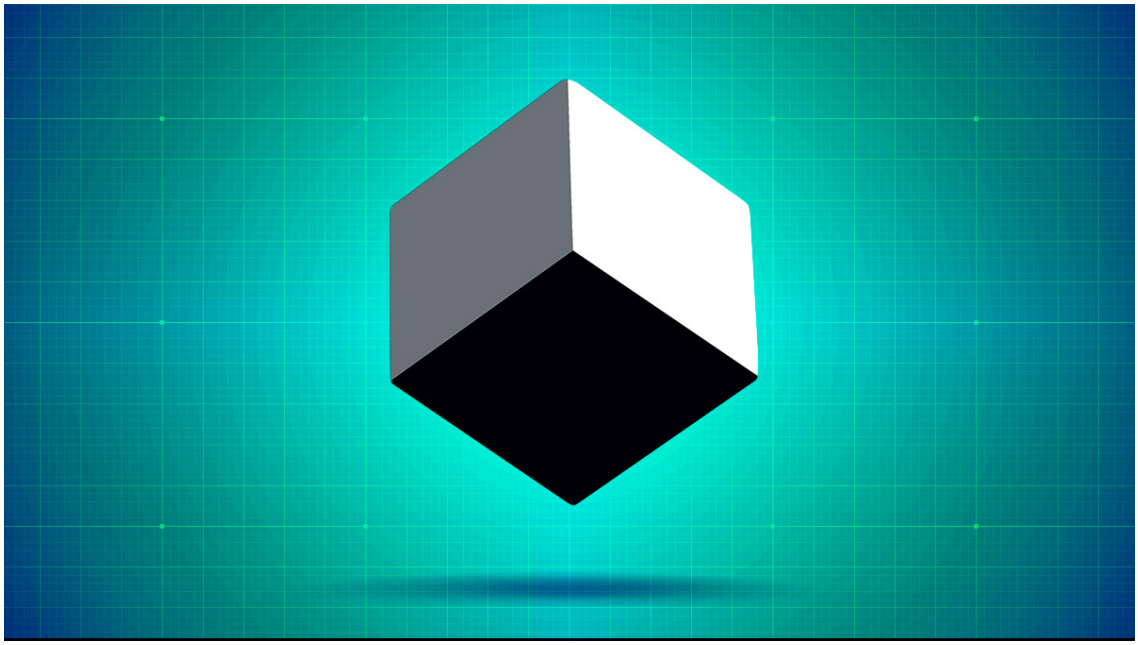
Harvard
Business
Review

AI And Machine Learning

When — and Why — You Should Explain How Your AI Works

by Reid Blackman and Beena Ammanath

August 31, 2022



HBR Staff/barleyman/Getty Images

Summary. AI adds value by identifying patterns so complex that they can defy human understanding. That can create a problem: AI can be a black box, which often renders us unable to answer crucial questions about its operations. That matters more in some cases than... **more**

"With the amount of data today, we know there is no way we as human beings can process it all...The only technique we know that can harvest insight from the data, is artificial intelligence,"

IBM CEO Arvind Krishna recently told the *Wall Street Journal*.

The insights to which Krishna is referring are patterns in the data that can help companies make predictions, whether that's the likelihood of someone defaulting on a mortgage, the probability of developing diabetes within the next two years, or whether a job candidate is a good fit. More specifically, AI identifies *mathematical* patterns found in thousands of variables and the relations among those variables. These patterns can be so complex that they can defy human understanding.

This can create a problem: While we understand the variables we put into the AI (mortgage applications, medical histories, resumes) and understand the outputs (approved for the loan, has diabetes, worthy of an interview), we might not understand what's going on between the inputs and the outputs. The AI can be a "black box," which often renders us unable to answer crucial questions about the operations of the "machine": Is it making reliable predictions? Is it making those predictions on solid or justified grounds? Will we know how to fix it if it breaks? Or more generally: can we trust a tool whose operations we don't understand, particularly when the stakes are high?

To the minds of many, the need to answer these questions leads to the demand for *explainable* AI: in short, AI whose predictions we can explain.

What Makes an Explanation Good?

A good explanation should be intelligible to its intended audience, and it should be useful, in the sense that it helps that audience achieve their goals. When it comes to explainable AI, there are a variety of stakeholders that might need to understand how an AI made a decision: regulators, end-users, data scientists, executives charged with protecting the organization's brand, and impacted consumers, to name a few. All of these groups have different skill sets, knowledge, and goals — an average citizen wouldn't likely understand a report intended for data scientists.

So, what counts as a good explanation depends on which stakeholders it's aimed at. Different audiences often require different explanations.

For instance, a consumer turned down by a bank for a mortgage would likely want to understand why they were denied so they can make changes in their lives in order to get a better decision next time. A doctor would want to understand why the prediction about the patient's illness was generated so they can determine whether the AI notices a pattern they do not or if the AI might be mistaken. Executives would want explanations that put them in a position to understand the ethical and reputational risks associated with the AI so they can create appropriate risk mitigation strategies or decide to make changes to their go to market strategy.

Tailoring an explanation to the audience and case at hand is easier said than done, however. It typically involves hard tradeoffs between accuracy and explainability. In general, reducing the complexity of the patterns an AI identifies makes it easier to understand how it produces the outputs it does. But, all else being equal, turning down the complexity can also mean turning down the accuracy — and thus the utility — of the AI. While data scientists have tools that offer insights into how different variables may be shaping outputs, these only offer a best guess as to what's going on inside the model, and are generally too technical for consumers, citizens, regulators, and executives to use them in making decisions.

Organizations should resolve this tension, or at least address it, in their approach to AI, including in their policies, design, and development of models they design in-hour or procure from third-party vendors. To do this, they should pay close attention to when explainability is a need to have versus a nice to have versus completely unnecessary.

When We Need Explainability

Attempting to explain how an AI creates its outputs takes time and resources; it isn't free. This means it's worthwhile to assess whether explainable outputs are needed in the first place for any particular use case. For instance, image recognition AI may be used to help clients tag photos of their dogs when they upload their photos to the cloud. In that case, accuracy may matter a great deal, but exactly *how* the model does it may not matter so much. Or take an AI that predicts when the shipment of screws will arrive at the toy factory; there may be no great need for explainability there. More generally, a good rule of thumb is that explainability is probably not a need-to-have when low risk predictions are made about entities that aren't people. (There are exceptions, however, as when optimizing routes for the subway leads to giving greater access to that resource to some subpopulations than others).

The corollary is that explainability may matter a great deal, especially when the outputs directly bear on how people are treated. There are at least four kinds of cases to consider in this regard.

When regulatory compliance calls for it.

Someone denied a loan or a mortgage deserves an explanation as to why they were denied. Not only do they deserve that explanation as a matter of respect — simply saying "no" to an applicant and then ignoring requests for an explanation is disrespectful — but it's also required by regulations. Financial services companies, which already require explanations for their non-AI models, will plausibly have to extend that requirement to AI models, as current and pending regulations, particularly out of the European Union, indicate.

When explainability is important so that end users can see how best to use the tool.

We don't need to know how the engine of a car works in order to drive it. But in some cases, knowing how a model works is imperative for its effective use. For instance, an AI that flags potential cases of fraud may be used by a fraud detection agent. If they do not know why the AI flagged the transaction, they won't know where to begin their investigation, resulting in a highly inefficient process. On the other hand, if the AI not only flags transactions as warranting further investigation but also comes with an explanation as to why the transaction was flagged, then the agent can do their work more efficiently and effectively.

When explainability could improve the system.

In some cases, data scientists can improve the accuracy of their models against relevant benchmarks by making tweaks to how it's trained or how it operates without having a deep understanding of how it works. This is the case with image recognition AI, for example. In other cases, knowing how the system works can help in debugging AI software and making other kinds of improvements. In those cases, devoting resources to explainability can be essential for the long-term business value of the model.

When explainability can help assess fairness.

Explainability comes, broadly, in two forms: global and local. Local explanations articulate why this particular input led to this particular output, for instance, why this particular person was denied a job interview. Global explanations articulate more generally how the model transforms inputs to outputs. Put differently, they articulate the rules of the model or the rules of the game. For example, people who have this kind of medical history with these kinds of blood test results get this kind of diagnosis.

In a wide variety of cases, we need to ask whether the outputs are fair: should this person really have been denied an interview or did we unfairly assess the candidate? Even more importantly, when we're asking someone to play by the rules of the hiring/mortgage lending/ad-receiving game, we need to assess whether the rules of the game are fair, reasonable, and generally ethically acceptable. Explanations, especially of the global variety, are thus important when we want or need to ethically assess the rules of the game; explanations enable us to see whether the rules are *justified*.

Building an Explainability Framework

Explainability matters in some cases and not in others, and when it does matter, it may matter for a variety of reasons. What's more, operational sensitivity to such matters can be crucial for the efficient, effective, and ethical design and deployment of AI. Organizations should thus create a framework that addresses the risks of black boxes to their industry and their organizations in particular, enabling them to properly prioritize explainability in each of their AI projects. That framework would not only enable data scientists to build models that work well, but also empower executives to make wise decisions about what should be designed and when systems are sufficiently trustworthy to deploy.

RB

Reid Blackman is the author of *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI* (Harvard Business Review Press, July 2022) and founder and CEO of Virtue, an ethical risk consultancy. He is also a senior adviser to the Deloitte AI Institute, previously served on Ernst & Young's AI Advisory Board, and volunteers as the chief ethics officer to the nonprofit Government Blockchain Association. Previously, Reid was a professor of philosophy