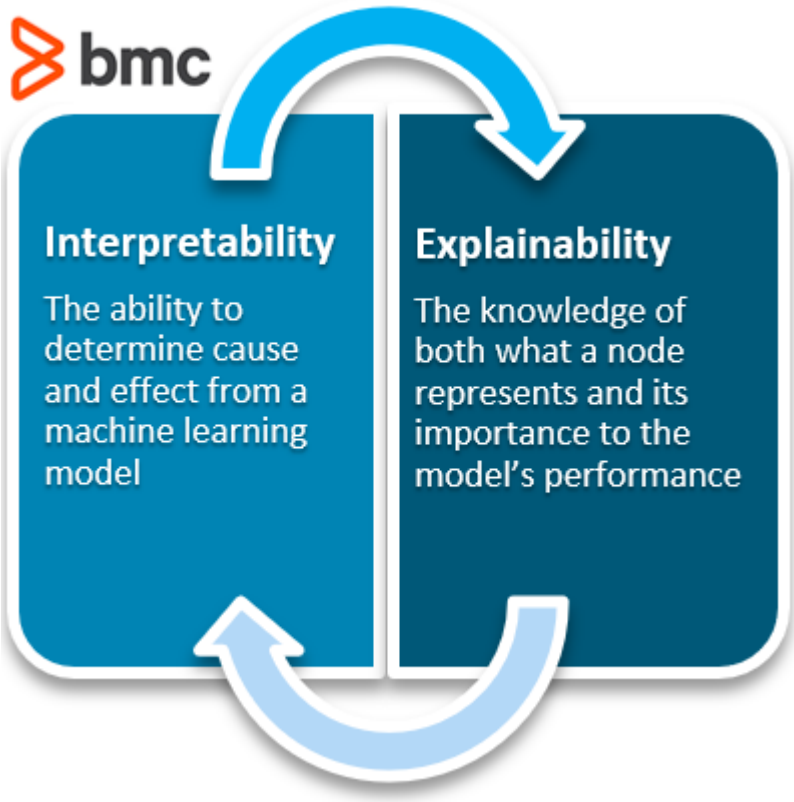
INTERPRETABILITY VS EXPLAINABILITY: THE BLACK BOX OF MACHINE LEARNING



Interpretability has to do with how accurate a machine learning model can associate a cause to an effect. Explainability has to do with the ability of the parameters, often hidden in Deep Nets, to justify the results.

This is a long article. Hang in there and, by the end, you will understand:

- How interpretability is different from explainability
- Why a model might need to be interpretable and/or explainable
- Who is working to solve the black box problem—and how



What is interpretability?

Does Chipotle make your stomach hurt? Does loud noise accelerate hearing loss? Are women less aggressive than men? If a machine learning model can create a definition around these relationships, it is interpretable.

All models must start with a hypothesis. Human curiosity propels a being to intuit that one thing relates to another. "Hmm...multiple black people shot by policemen...seemingly out of proportion to other races...something might be systemic?" Explore.

People create internal models **to interpret** their surroundings. In the field of machine learning, these models can be tested and verified as either accurate or inaccurate representations of the world.

Interpretability means that the cause and effect can be determined.

If a model can take the inputs, and routinely get the same outputs, the model is interpretable:

- If you overeat your pasta at dinnertime and you always have troubles sleeping, the situation is interpretable.
- If all 2016 polls showed a Democratic win and the Republican candidate took office, all those models showed low interpretability. If the pollsters' goal is to have a good model, which the institution of journalism is compelled to do—report the truth—then the error shows their models need to be updated.

Low interpretability

Interpretability poses no issue in low-risk scenarios. If a model is recommending movies to watch, that can be a low-risk task. (Unless you're one of the big content providers, and all your recommendations suck to the point people feel they're wasting their time, but you get the picture). If

a model is generating what color will be your favorite color of the day or generating simple yogi goals for you to focus on throughout the day, they play low-stakes games and the interpretability of the model is unnecessary.

The necessity of high interpretability

Interpretability sometimes needs to be high in order to justify why one model is better than another.

In *Moneyball*, the old school scouts had an interpretable model they used to pick good players for baseball teams; these weren't machine learning models, but the scouts had developed their methods (an algorithm, basically) for selecting which player would perform well one season versus another. But the head coach wanted to change this method.

For Billy Beane's methods to work, and for the methodology to catch on, his model had to be highly interpretable when it went against everything the industry had believed to be true. High model interpretability wins arguments.

Risk and responsibility

A model with high interpretability is desirable on a high-risk stakes game. High interpretable models equate to being able to hold another party liable. And when models are predicting whether a person has cancer, people need to be held accountable for the decision that was made. Highly interpretable models, and maintaining high interpretability as a design standard, can help build trust between engineers and users.

It's bad enough when the chain of command prevents a person from being able to speak to the party responsible for making the decision. It is much worse when there is no party responsible and it is a machine learning model to which everyone pins the responsibility. There is no retribution in giving the model a penalty for its actions.

When Theranos failed to produce accurate results from a "single drop of blood", people could back away from supporting the company and watch it and its fraudulent leaders go bankrupt.

Finally, high interpretability allows people to play the system. If the teacher hands out a rubric that shows how they are grading the test, all the student needs to do is to play their answers to the test. If the teacher is a *Wayne's World* fanatic, the student knows to drop anecdotes to Wayne's World. Or, if the teacher really wants to make sure the student understands *the process* of how bacteria breaks down proteins in the stomach, then the student shouldn't describe the kinds of proteins and bacteria that exist. Instead, they should jump straight into what the bacteria is doing.

Students figured out that the automatic grading system or the SAT couldn't actually comprehend what was written on their exams. Somehow the students got access to the information of a highly interpretable model. The ML classifiers on the Robo-Graders scored longer words higher than shorter words; it was as simple as that. A string of 10-dollar words could score higher than a complete sentence with 5-cent words and a subject and predicate.

As VICE reported, "'The BABEL Generator proved you can have complete incoherence, meaning one sentence had nothing to do with another,' and still receive a high mark from the algorithms." Of course, students took advantage.

Having worked in the NLP field myself, these still aren't without their faults, but people are creating

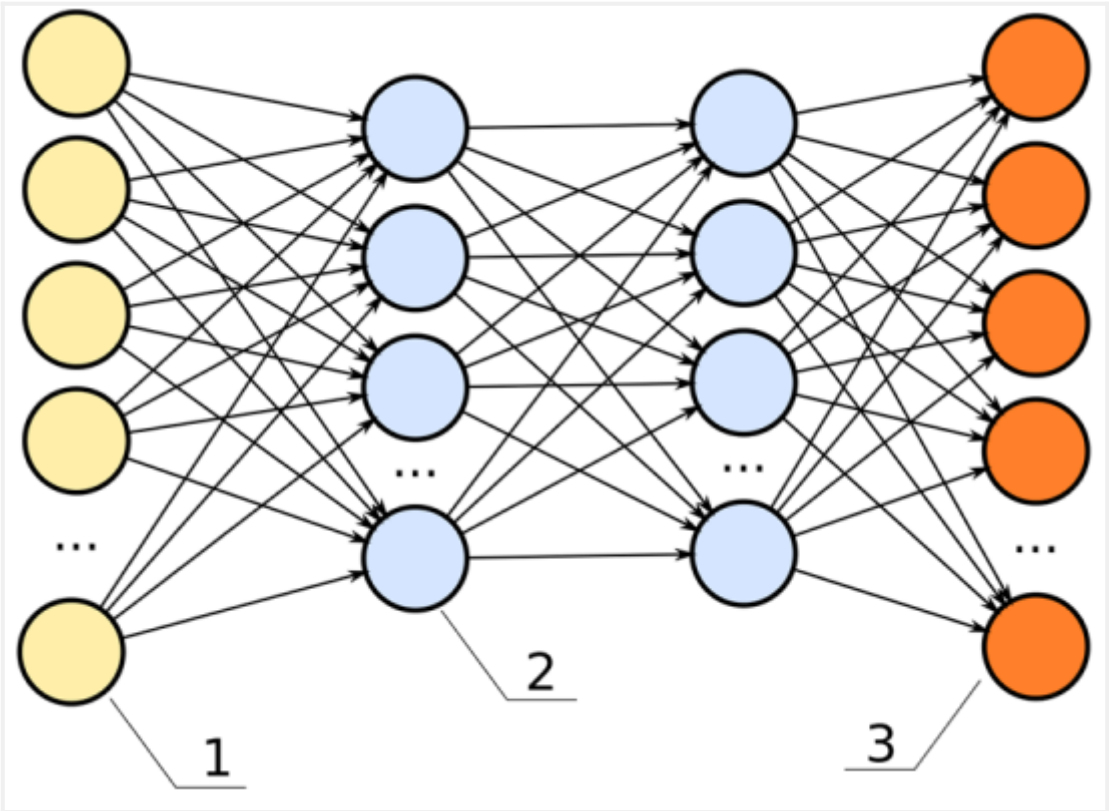
ways for the algorithm to know when a piece of writing is just gibberish or if it is something at least moderately coherent.

What is explainability?

ML models are often called black-box models because they allow a pre-set number of empty parameters, or nodes, to be assigned values by the machine learning algorithm. Specifically, the back-propagation step is responsible for updating the weights based on its error function.

To predict when a person might die—the fun gamble one might play when calculating a life insurance premium, and the strange bet a person makes against their own life when purchasing a life insurance package—a model will take in its inputs, and output a percent chance the given person has at living to age 80.

Below is an image of a neural network. The inputs are the yellow; the outputs are the orange. Like a rubric to an overall grade, explainability shows how significant each of the parameters, all the blue nodes, contribute to the final decision.



In this neural network, the hidden layers (the two columns of blue dots) would be the black box.

For example, we have these data inputs:

- Age
- BMI score
- Number of years spent smoking
- Career category

If this model had high explainability, we'd be able to say, for instance:

- The career category is about 40% important

- The number of years spent smoking weighs in at 35% important
- The age is 15% important
- The BMI score is 10% important

Explainability: important, not always necessary

Explainability becomes significant in the field of machine learning because, often, it is not apparent. Explainability is often unnecessary. A machine learning engineer can build a model without ever having considered the model's explainability. It is an extra step in the building process—like wearing a seat belt while driving a car. It is unnecessary for the car to perform, but offers insurance when things crash.

The benefit a deep neural net offers to engineers is it creates a black box of parameters, like fake additional data points, that allow a model to base its decisions against. These fake data points go unknown to the engineer. The black box, or hidden layers, allow a model to make associations among the given data points to predict better results. For example, if we are deciding how long someone might have to live, and we use career data as an input, it is possible the model sorts the careers into high- and low-risk career options all on its own.

Perhaps we inspect a node and see it relates oil rig workers, underwater welders, and boat cooks to each other. It is possible the neural net makes connections between the lifespan of these individuals and puts a placeholder in the deep net to associate these. If we were to examine the individual nodes in the black box, we could note this clustering interprets water careers to be a high-risk job.

In the previous chart, each one of the lines connecting from the yellow dot to the blue dot can represent a signal, weighing the importance of that node in determining the overall score of the output.

- If that signal is high, that node is significant to the model's overall performance.
- If that signal is low, the node is insignificant.

With this understanding, we can define explainability as:

> *Knowledge of what one node represents and how important it is to the model's performance.*

Questioning the "how"?

Image classification tasks are interesting because, usually, the only data provided is a sequence of pixels and labels of the image data. The general purpose of using image data is to detect what objects are in the image. If you were to input an image of a dog, then the output should be "dog". *How* this happens can be completely unknown, and, as long as the model works (high interpretability), there is often no question as to how.

In image detection algorithms, usually Convolutional Neural Networks, their first layers will contain references to shading and edge detection. The human never had to explicitly define an edge or a shadow, but because both are common among every photo, the features cluster as a single node and the algorithm ranks the node as significant to predicting the final result. The image detection model becomes more explainable.

Google apologized recently for the results of their model. As the headline likes to say, their algorithm produced racist results. Machine learning models are not generally used to make a single

decision. They're created, like software and computers, to make many decisions over and over and over. Machine learning models are meant to make decisions at scale. If those decisions happen to contain biases towards one race or one sex, and influence the way those groups of people behave, then it can err in a very big way.

The equivalent would be telling one kid they can have the candy while telling the other they can't. The decision will condition the kid to make behavioral decisions without candy. With ML, this happens at scale and to everyone. And—a crucial point—most of the time, the people who are affected have no reference point to make claims of bias. They just know something is happening they don't quite understand.

In support of explainability

For the activist enthusiasts, explainability is important for ML engineers to use in order to ensure their models are not making decisions based on sex or race or any other data point they wish to make ambiguous. The decisions models make based on these items can be severe or erroneous from model-to-model.

Fortunately, in a free, democratic society, there are people, like the activists and journalists in the world, who keep companies in check and try to point out these errors, like Google's, before any harm is done. In a society with independent contractors and many remote workers, corporations don't have dictator-like rule to build bad models and deploy them into practice. The workers at many companies have an easier time reporting their findings to others, and, even more pivotal, are in a position to correct any mistakes that might slip while they're hacking away at their daily grind.

Good communication, and democratic rule, ensure a society that is self-correcting. It is persistently true in resilient engineering and chaos engineering. It is true when avoiding the corporate death spiral. It is a reason to support explainable models. Explainable models (XAI) improve communication around decisions.

By exploring the explainable components of a ML model, and tweaking those components, it is possible to adjust the overall prediction. To point out another hot topic on a different spectrum, Google had a competition appear on Kaggle in 2019 to "end gender bias in pronoun resolution". Basically, natural language processes (NLP) uses use a technique called coreference resolution to link pronouns to their nouns. It's become a machine learning task to predict the pronoun "her" after the word "Shauna" is used.

Shauna likes racing. It's her favorite sport.

Coreference resolution will map:

- Shauna her
- racing it

This technique can increase the known information in a dataset by 3-5 times by replacing all unknown entities—the shes, his, its, theirs, thems—with the actual entity they refer to— Jessica, Sam, toys, Bieber International. The goal of the competition was to uncover the internal mechanism that explains gender and reverse engineer it to turn it off.

It might be thought that big companies are not fighting to end these issues, but their engineers are actively coming together to consider the issues. Economically, it increases their goodwill.

I used Google quite a bit in this article, and Google is not a single mind. Google is a small city, sitting at about 200,000 employees, with almost just as many temp workers, and its influence is incalculable. Amazon is at 900,000 employees in, probably, a similar situation with temps. That is far too many people for there to exist much secrecy. Enron sat at 29,000 people in its day.

Solving the black box problem

Finally, to end with Google on a high, Susan Ruyu Qi put together an article with a good argument for why Google DeepMind might have fixed the black-box problem. The point is: explainability is a core problem the ML field is actively solving. With everyone tackling many sides of the same problem, it's going to be hard for something really bad to slip under someone's nose undetected.

Computers have always attracted the outsiders of society, the people whom large systems always work against. These people look in the mirror at anomalies every day; they are the perfect watchdogs to be polishing lines of code that dictate who gets treated how. Explainability and interpretability add an observable component to the ML models, enabling the watchdogs to do what they are already doing.

There's also promise in the new generation of 20-somethings who have grown to appreciate the value of the whistleblower. They maintain an independent moral code that comes before all else. If you don't believe me: Why else do you think they hop job-to-job?

Additional resources

Explore the BMC Machine Learning & Big Data Blog and these related resources:

- Machine Learning: Hype vs Reality
- Enabling the Citizen Data Scientists
- Top 5 Machine Learning Algorithms for Beginners
- Multi-part Guides with tutorials:
    - Apache Spark & Machine Learning
    - Machine Learning with sciKit Learn
    - Machine Learning with TensorFlow and Keras