


AI Makes Decisions We Don't Understand. That's a Problem.

Fixing it won't be easy, however.

 AI Makes Decisions We Don't Understand. That's a Problem.

[Machine learning](#) algorithms, the technology that powers AI, have advanced quickly in recent decades. Today, [deep learning](#) algorithms power facial recognition software and enable anyone to create realistic [deepfake](#) photos and videos in just a few minutes.

AI algorithms are also increasingly used by companies and institutions, from creating smart voice assistants to generating [automatic language translations](#). But along with the growing adoption of AI is the problem that AI models are not well understood — much of the time, people don't know why AI models make certain determinations.

And it's not just the average person off the street who doesn't understand — even the researchers and programmers creating them don't really understand why the models they have built make the decisions they make.

“If you cannot adequately describe what your algorithm is doing, that seems problematic,” said Bryan Routledge, associate professor of finance at Carnegie Mellon University.

Not being well understood by its own creator is a strange phenomenon of AI, but it's also the reason for its power and success — using AI methods, people can create something that's self-training and able to perform certain calculations beyond people's capabilities.

“If you cannot adequately describe what your algorithm is doing, that seems problematic.”

This phenomenon has created a growing gap between what we're able to do with AI techniques and our ability as humans to understand the technology we use, and it's part of the reason some have called for a “right to explanation” when it comes to the use of AI.

Routledge and co-author Tae Wan Kim recently [published an analysis](#) on the right to explanation, concluding that the public has an ethical right to know how companies' AI models make decisions. They reasoned that consumers should be able to demand explanations for specific AI decisions that appear biased, and also to learn about how those models make decisions so they can make informed choices about which companies to give their business to.

Routledge said AI models should be held to a similar standard as medicine, another complex field that makes important decisions about people, but whose decisions are backed up by explanations.

“Being able to explain things in a way that patients are informed is part of what it means to be a doctor,” he said. “Medicine has been dealing with this for a long time, so it's fairly second nature. When there's a new treatment, part of what they think about is, ‘How will we properly communicate this to patients so that they can be informed when they choose yes or no?’”

More on Explainable AI [Weighing the Trade-Offs of Explainable AI](#)

Explanations Are Important When Humans Can't Verify AI's Work

Not every use case for AI is equally in need of explanations, said Rayid Ghani, a machine learning professor at Carnegie Mellon University. For example, it's usually easy to tell if an image recognition program has been mislabeling things — like if an image labeled “dog” actually depicts a cat. And some use cases are not consequential enough to require explanations.

“If you're showing somebody an ad, I don't really care if you have a great explanation or not,” he said. “The cost to me of seeing a good or a bad ad, there's very little difference — I don't care, I don't want to see any ads.”

Ghani said AI technology is used to accomplish two categories of tasks: either tasks humans are good at doing but do slowly, or tasks humans are incapable of doing. For instance, humans are great at recognizing and labeling images. AI models can be trained to do that as well — not necessarily better than humans, but much, much faster.

On the other hand, AI models can also be used to help make predictions — something people do not do well.

“Who's going to be involved in a shooting, or who's going to not graduate from high school on time, or who's going to be unemployed long term — humans are not very good at that problem,” Ghani said.

Understanding how AI models make their decisions is particularly important for those types of predictive use cases because humans aren't able to verify whether the model is working correctly. Unlike image categorization, where humans can simply look at the images to check whether they have been labeled correctly, predictive AI gives outputs that humans can't verify for themselves.

“It sort of helps you figure out how to separate the correct predictions from the incorrect predictions.”

“So when an explanation comes from the system, we don't really know whether it's correct or not because we don't have that understanding. If we did, we wouldn't need a computer to help us solve that problem,” Ghani said. “Because we're not trying to be efficient — we're trying to be better than humans.”

Ghani welcomed the idea of a right to explanation, anticipating many benefits once it's possible to explain the decisions AI models make. Explanations could help people figure out how to change their behavior to get better results and detect if models are drawing conclusions based on faulty reasoning.

Ghani gave the example of a hypothetical medical AI model that predicts patient outcomes. If explanations of the model revealed that a significant factor behind its decision making was based on the day of the week patients are admitted, people would know to be suspicious of the accuracy of the model — or the hospital would know to investigate this unacceptable pattern in patient prognosis.

“It sort of helps you figure out how to separate the correct predictions from the incorrect predictions,” Ghani said. “It helps you sanity check.”

Perhaps most importantly, explanations might be able to satisfy that part of us that is rightfully suspicious when asked to trust things we don't understand, and to answer our many questions: Why did the AI model make that particular prediction? Is the AI model actually answering the question we're asking in the way that we expect? Is it reliable?

What Do We Mean by ‘Explanation’?

How did we even get here, where even the experts creating AI models don’t understand how those models are making decisions?

The way machine learning works is it uses a lot of data to refine models. The data scientist takes an AI algorithm, points it at a desired result and basically lets the algorithm run itself using some initial random values and a mountain of testing data. Eventually, the algorithm will refine the model into something that can achieve the desired result. AI algorithms take advantage of computers’ abilities to do math and calculate at great speeds.

But the outcome from this brand of problem-solving is that the interactions in the model built by absorbing so much data are too much for a person to wrap their minds around. That’s what makes many AI models black boxes — difficult or impossible to understand.

Because the way AI models work is so specific and math-based, as you dig into the topic of explainable AI, you eventually run into the fundamental question of what basic ideas like “explanation” and “understanding” really mean.

“Who decides if something is explainable? If I gave you an explanation, is that a good explanation or a bad one?”

It may seem like a silly debate over semantics, but how these concepts are defined could have real-world impacts on business and regulatory decisions if a right to explanation comes to pass. That’s because good regulation first requires precise and accurate definitions.

“‘Explainable’ and ‘accurate’ — in some ways, both of those are ambiguous concepts,” Ghani said. “Who decides if something is explainable? If I gave you an explanation, is that a good explanation or a bad one? What does it mean to be good? What does it mean to be bad?”

There’s one way of defining “explanation” that is quite simple, but doesn’t satisfy our human search for understanding: just follow the data as it flows through the model’s mathematical functions, tracing through all the paths until it finally becomes the model’s output.

That definition is rooted in the idea that algorithms are like machines — the only pure way to “understand” one is to follow the machine’s mechanism. “Why” is a human question that doesn’t apply to machines. You would never ask a car “why” — the real question is “how.” That same logic should be applied to algorithms, argues Zachary Lipton, who runs the [Approximately Correct Machine Intelligence Lab](#) at Carnegie Mellon University.

Find out who's **hiring**.

See all **Developer + Engineer** jobs at top tech companies & startups

View 10000+ Jobs

“One question you might ask is, ‘Account to me how the models arrived at the answer?’” Lipton said. “And the answer is: There is no specific answer besides the weights of the model.”

Still, that might provide some understanding for very simple models. But AI models created by algorithms like deep learning can easily take in data with thousands of attributes. You can trace inputs

through those models, but it would not provide a deep-level understanding of how the model arrives at its conclusions — mostly because of human limitations.

“A deep neural network that was trained on images takes every single possible input in a 400,000-dimensional space,” Lipton said. “The full mapping is not something that you can put in your head.”

Who the audience is that’s getting an explanation can also affect how explanations should be defined because people have different levels of understanding when it comes to AI concepts. Explaining a model to an AI researcher may be very different from explaining it to a layperson, policymaker or business person.

“A deep neural network that was trained on images takes every single possible input in a 400,000-dimensional space. The full mapping is not something that you can put in your head.”

When you dig into it, it’s hard to pin down exactly what it means to make AI explainable. People are used to getting explanations from other people, such as when someone gives the reasoning behind a decision they made or an action they took. But how do we know the reasons they give are an accurate account of their true motivations? How do we know theirs is a full explanation?

When people take actions and make decisions, they’re not drawing just from pure logic, but also from personal experience, knowledge, emotion and their personalities. Most of the time, someone’s explanation is probably more of an approximation, the very top layer of a jumble of subconscious factors. Maybe they don’t even know the true motivation behind their behavior themselves.

AI models built using deep learning algorithms and large amounts of data also take on some narrow form of this complexity, mostly in terms of experience and knowledge.

So which layer of explanation do we want? Maybe what people want is for the way AI models actually reach decisions to be approximated and simplified enough so that we can wrap our minds around the whole process all at once. But is it possible to have that and still call it an explanation?

The (Questionable) Promise of Explainable AI

There is already a field of study known as [explainable AI](#). Its researchers use mathematical techniques to examine patterns in AI models and draw conclusions about how those models reach their decisions. Many explainable AI techniques are “general” techniques, meant to be applicable for explaining any kind of machine learning model.

But Lipton considers the current field of explainable AI to be littered with false promises. He said the benefits of applying explainable AI techniques to any type of AI model, which makes these techniques compelling, also makes them incapable of explaining anything meaningful.

Some explainable AI techniques for labeling images, for example, black out sections of an image at a time. They then run the altered images through the original AI model to see what differences the new output has from the original.

But Lipton hypothesized that blacking out parts of images may render them so unlike natural images that it affects models’ outcomes beyond what researchers might expect. Although explainable AI methods use math to get their results, he said, those mathematical techniques have not yet been proven to offer insights into AI models.

“Many of these methods are providing something that really doesn’t tell you anything at all.”

“If there’s two equations in what you’re calling an explanation, and you’re presenting it to someone who’s mathematically illiterate, they don’t have the ability to look at that and call bullshit,” Lipton said. “Many of these methods are providing something that really doesn’t tell you anything at all.”

He said AI researchers can change explainable AI conclusions drastically by making small tweaks to AI models while retaining the same model outcomes.

“Clearly, it wasn’t an explanation in the first place, if you’re able to induce whatever explanations you want without changing what you’re actually doing,” he said.

The ambiguity of the term “explanation” is part of what concerns Lipton. Because the definition of explanation for AI models is so loosely defined, people may come to accept math-heavy “explanations” that confuse and dazzle, but don’t actually provide real answers.

“A huge fraction of the field is overrun by precisely that sort of thing,” Lipton said. “Basically what people do is they propose some algorithm and it’s just some sort of trick.”

It can also be difficult to understand the explanations offered. If an explanation of an image classifier simply [highlights areas on an image](#), is that really an explanation? What is it really saying about how the model makes decisions?

“The problem is they’re being fooled by a cottage industry that thinks you can somehow take predictive machine learning models, kind of wave a magic wand on them, generate some images, and call that an explanation.”

It’s even possible to get into a scenario where you need another algorithm to interpret the explanation, Ghani said.

“It’s sort of this weird thing where now you have to build another system, on top of the explanation system, to sort of explain the explainer,” he said.

Though Lipton is opposed to the methods used by the current field of explainable AI, he is sympathetic to the right to explanation’s core goal of being able to understand what AI models are doing.

“People think that certain decisions should be guided by sound reasoning and you should be able to provide explanations when you make certain categories of decisions — and those people are absolutely right,” Lipton said. “The problem is they’re being fooled by a cottage industry that thinks you can somehow take predictive machine learning models, kind of wave a magic wand on them, generate some images, and call that an explanation.”

He worries that people accepting anything as an “explanation” would instead be enabling unethical uses of AI.

“The worst-case scenario is that you have the academic community be complicit in giving people the impression that they actually now have an academic seal of approval,” Lipton said.

More on Machine Learning [Inside the Machine Learning Effort to Organize the Library of Congress Digital Collection](#)

Ask Specific Questions Instead

Instead of trying to explain an entire AI model all at once, it may be more effective to analyze models using tools specific to the questions you want to ask and to the AI model.

That's because AI algorithms and AI models vary widely. For instance, there's not only one way of doing deep learning — it's actually a whole category of methods, and each use case of a method can be wildly different depending on the training data and what researchers are optimizing for.

"It's not like you download [a] deep learning [model] and you click a button and run it," Ghani said. "When you build these types of systems, you build hundreds of versions of them. ... You then define a performance metric that you care about, and you choose the one that's doing well on that performance metric."

Machine learning models output results in the form of probabilities rather than clear-cut yes and no answers. An image recognition program would predict that one image has a 40 percent chance of being a cat and another has an 89 percent likelihood, instead of saying this one isn't a cat and this one is. It's up to data scientists to determine what the cutoff for labeling should be.

"We need to be very deliberate in designing systems that are focused on a specific use case and a specific user in designing an explanation."

If it's more important that the model not miss any images with cats — even if a few non-cat images are mistaken for cats — the labeling threshold for cats should be set at a low percentage. On the other hand, if it's expensive to review the image labels, the maker might want to set a high threshold. Each of these variations corresponds to a different AI model.

"In the explainable AI world, we're trying to build this general-purpose, monolithic explainable AI system that's supposed to just do everything for every type of problem, for every type of model, for every type of user — and I think that's just wrong," Ghani said. "We need to be very deliberate in designing systems that are focused on a specific use case and a specific user in designing an explanation, and then experimentally validating whether that explanation helps that user achieve the goal of the system."

Some statistical methods already exist that are effective at answering very specific questions, such as ablation tests and counterfactuals, Lipton said. If a bank used AI models to determine whether to approve mortgages for clients and turned someone down, counterfactuals could give people concrete feedback on what they can improve to change the result next time.

"If they said, 'You have four lines of credit and a late payment in the last year. But if you had six lines of credit and no late payment for at least 18 months, then you would have been approved,'" Lipton said. "So this is affording some degree of transparency and providing someone with something actionable."

More Investigation Is Needed

The only thing certain about explainability in AI is that there's still plenty of room left for investigation.

My first introduction to explainability in AI was talking to a professor about the [use of AI in the insurance industry](#). He cautioned that AI would make insurance companies difficult to audit because questions of bias can't be evaluated if the insurance companies themselves don't even know how insurance decisions are made.

Curiously, both Lipton and Ghani pushed back against the idea of using explanations in AI to help determine bias in AI models. They argued that the two concepts are not related because explaining why an AI model produced a given output doesn't provide any insight into whether the overall model is biased.

That's partly why some who oppose a right to explanation argue that monitoring AI models' results over time for hints of bias, and adjusting when needed, is better than requiring explanations from AI models.

Routledge, co-author of the right to explanation analysis, said monitoring AI model results for bias over time is a good practice, but not a substitute.

"If you applied for a loan and were denied, while someone who looks similar to you got a loan, and you ask why that is, and the company's answer is, 'Don't worry, we adjust things as we go forward' — that's not very satisfying," he said. "It doesn't seem like it would be an adequate explanation."

There are plenty of people who oppose explaining AI models at all. One common argument is that it limits AI's potential by tying it down to unnecessary human constraints. Another is that a right to explanation would be impossible to enforce, especially considering how imprecise the human concept of "understanding" is.

"When you actually have real concerns about discrimination ... maybe what should actually be happening is people should be taking the technology off the table altogether."

Lipton himself favors banning the use of AI outright for certain cases instead of using unreliable techniques to explain AI models and spending time debating the definition of "explain."

"When you actually have real concerns about discrimination ... maybe what should actually be happening is people should be taking the technology off the table altogether," Lipton said. "Like, 'Algorithmic resume screening is not a mature or appropriate technology and should not be applied.'"

Whether establishing a right to explanation is viable depends in large part on whether it's possible to develop techniques that can explain AI models in the first place, but there hasn't been nearly enough rigorous study in this field to figure out whether that's the case.

Ghani said a big hurdle for the field currently is using more realistic scenarios in research, which is important because AI models are built to perform specific tasks.

"One of the things that we're working on is building these collaborations with organizations — whether it's governments or nonprofits or companies — where things can be anchored in real problems, using real data, tested on real users, so that you can see what actually works or not," he said.

There are some reasons for cautious optimism. When people talk about black box AI models, such as those built by deep learning algorithms, there seems to be an implication that the more powerful and successful the algorithm, the more of a black box it is. But that's not necessarily the case.

"That concept is often used that there's this trade-off, but we actually don't know if there is a trade-off — it's not a universal concept that needs to be true," Ghani said. "I think people say that a lot, but I think people don't say it based on empirical evidence."

Maybe researchers can find ways to use existing techniques to explain AI models, or maybe new AI techniques can be developed that prioritizes explainability. Routledge, for one, was hopeful.

“I guess everything I’m saying is that it’s not easy, but it does not seem impossible,” he said.