Should AI Models Be Explainable? That depends.

A Stanford researcher advocates for clarity about the different types of interpretability and the contexts in which it is useful.



Artificial intelligence systems that can diagnose cancer, read medical images, or suggest an appropriate medication often get a bum rap for being unable to explain themselves: When probed, they offer up only a black box as to how they achieve their success. As a result, some clinicians wonder whether these AI tools should be trusted.

But AI models do not need to be interpretable to be useful, says Nigam Shah, professor of medicine (biomedical informatics) and of biomedical data

science at Stanford University and an affiliated faculty member of the Stanford Institute for Human-Centered Artificial Intelligence. That's especially true in medicine, he says, where doctors routinely offer treatments without knowing how or why they work.

"Of the 4,900 drugs prescribed on a routine basis, we don't fully know how most of them really work," says Shah. "But we still use them because we have convinced ourselves via randomized control trials that they are beneficial."

The same can be true of AI models. "Judicious testing should be enough," Shah says. If an AI model yields accurate predictions that help clinicians better treat their patients, then it may be useful even without a detailed explanation of how or why it works.

It's like the weather report, Shah says. "Do you, as a user, care how the weather is predicted, and what the causal explanation is, as long as you know a day ahead if it is going to rain and the forecast is correct?"

That's not to say that AI interpretability isn't valuable. Indeed, in contexts where AI models are used in an automated fashion to deny people job interviews, bail, loans, health care programs or housing, Shah says laws and regulations should absolutely require a *causal* explanation of these decisions to ensure that they are fair.

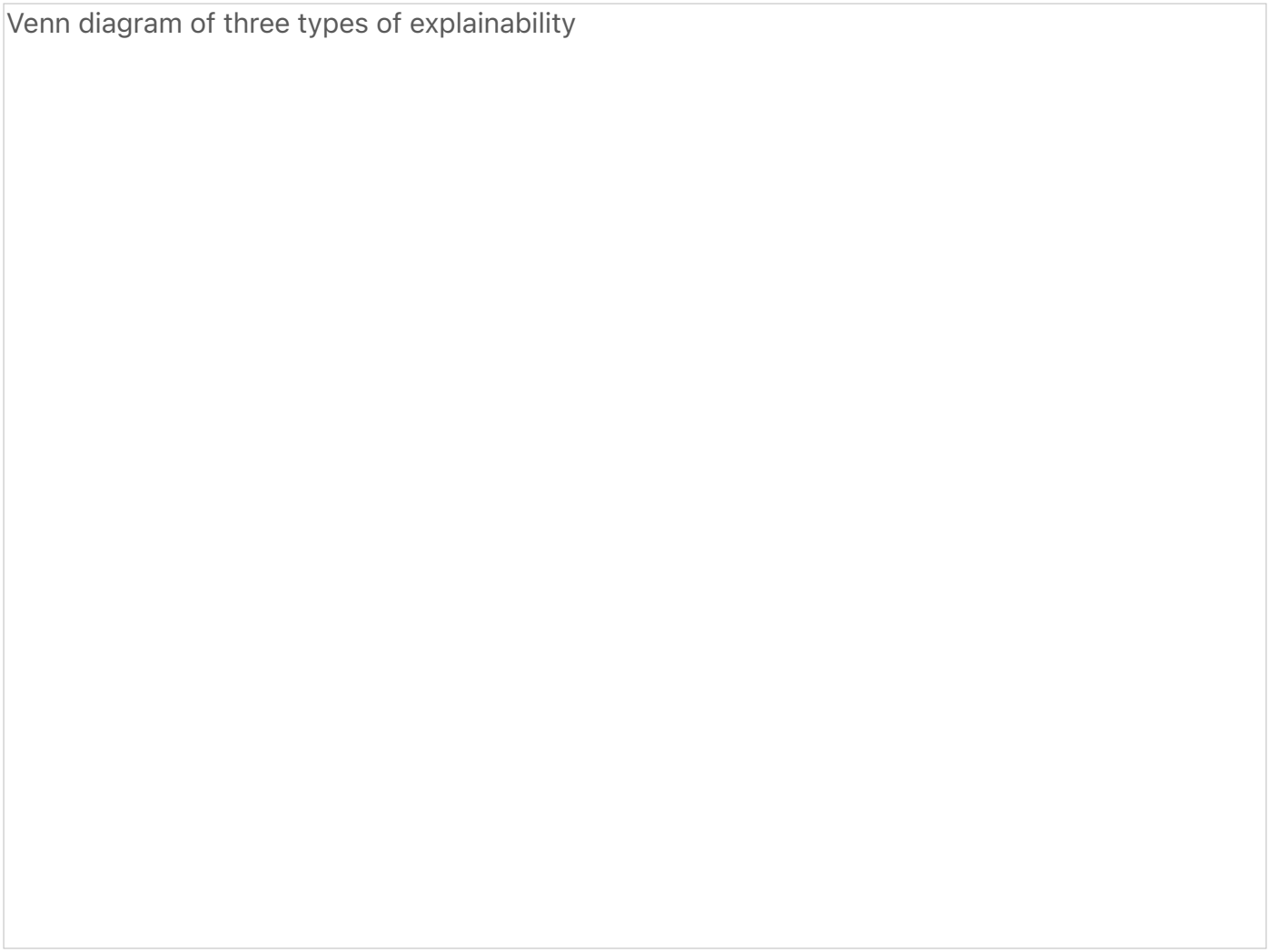
But in health care, where AI models rarely lead to such automated decision making, an explanation may or may not be useful, Shah says. "It is essential that model developers be clear about why an explanation is needed and what type of explanation is useful for a given situation."

Parsing Interpretability

Shah identifies three main types of AI interpretability: The engineers' version of explainability, which is geared toward *how* a model works; causal

explainability, which relates to *why* the model input yielded the model output; and trust-inducing explainability that provides the information people need in order to *trust* a model and confidently deploy it.

When researchers or clinicians are talking about interpretability, he says, "We need to know which of those we're aspiring for."



Some explanations might lie at the intersection of all three types (see diagram). One example is the commonly used atherosclerotic cardiovascular disease (ASCVD) risk equation, which relies on nine primary data points (including age, gender, race, total cholesterol, LDL/HDL-cholesterol, blood pressure, smoking history, diabetic status, and use of antihypertensive medications) to calculate a patient's 10-year risk of having a heart attack or stroke. The engineers' explanation of *how* the ASCVD risk calculator produces a prediction is simple: "All that's happening inside this algorithm is

multiplication and addition," Shah says. And the causal explanation for the ASCVD risk score is straightforward as well, because cholesterol levels and blood pressure are causally related to the risk of heart attack. Finally, the ASCVD risk equation is trusted and useful in clinical settings because the risk estimate and the causal explanation also point to an action that the clinician can take, e.g., prescribe cholesterol-lowering medication.

But step outside the bounds of this simple example and one cannot presume that the three types of explanations will overlap.

Indeed, engineers' explanations of complex AI models can reveal factors that legitimately contribute to a model's predictive power but have no causal relationship to the model's output and are unlikely to induce trust in the model. In one classic example published in *Nature*, a model for predicting hospital mortality found that factors such as visitation by a priest or assigning consent to next of kin were highly indicative of impending mortality. Clinicians, who intuitively want to know *why* a patient is going to die, might needlessly distrust such tautological predictions even though they are accurate. "If they ignore the model's high predictive accuracy and stop using the model to trigger the important next step of calling the family, that would be a mistake," Shah says. Instead of being helpful, the explanation has eroded the model's usefulness.

This scenario remains true even when a neural net provides a causal explanation for a mortality prediction – that is, when the prediction is based not on a priest's visit but on the symptoms that indicate why a patient will not last another 24 hours. But he says these explanations are useless to clinicians because it's really too late to intervene medically. "What you need to do is call the family," Shah says. "And for that, you don't need the whole fancy explanation. All you needed to know is that the last nine out of ten times this algorithm predicted 24-hour mortality, it was right."

The point, Shah says, is that an explanation can land in any of the seven zones of the Venn diagram above. Specific individuals may only care about

explanations from a specific zone. Engineers want to know how their model works in order to debug it, and clinicians may want to know a causal explanation in order to trust it and appropriately treat a patient. "Showing someone an explanation from a *different* zone than the one they need can have unexpected consequences, such as loss of trust in an accurate and useful model," Shah says.

The Trust Paradox

Explanations aren't always necessary. But even more concerning, sometimes they lead people to rely on a model even when it's wrong.

In some elegant experiments, Microsoft Research showed that people were more apt to accept obvious errors in an interpretable model due to a false sense of trust.

In their experiments, Microsoft Research asked people to predict the price of a New York City apartment given eight features, ranging from the number of bedrooms and bathrooms to the distance to subways or schools. Before making their predictions, participants were shown the rental price predicted by a statistical model. For some participants, the model's calculations were displayed. For others, the model was a black box: The participants didn't know how the model came up with its predictions. Interestingly, when the model was applied to an unusual apartment (e.g., one with more bathrooms than bedrooms), the participants who were shown the explained model were more likely to defer to the predictions than those who were shown the black box. "The people using the interpretable model kind of believed the output, but the people using the black box model saw it as suspicious," Shah says. The upshot: Maybe explanations sometimes increase trust too much. In fact, black boxes could be useful for keeping people's guard up.

Shah argues that when interpretability is invoked, it's important to tease out why it is needed before offering any type of explanation. It matters, he says,

"because if you offer an explanation, as in the apartment example, it might actually backfire."

More Acceptance for Unexplained AI?

For about four years, Shah has been arguing for greater clarity around interpretability and says he has been in the minority. But that's now changing. *Science* published a paper in 2019 titled "In defense of the black box," and Microsoft Research's apartment study challenged whether interpretability is helpful. Now, Shah says, some professors of computer science give talks titled "Don't explain yourself," suggesting that the engineers' version of interpretability might not meet the needs of the model users who care more about causality and trust. Indeed, explanations of *how* models work might just distract people from figuring out what they really need or want to know.

"The type of interpretability needed varies depending on the context," Shah says. "We should not insist on just one kind; and most importantly we should not confuse the different notions of interpretability because each kind serves a different purpose."

Stanford HAI's mission is to advance AI research, education, policy and practice to improve the human condition. Learn more.