Do AI Models Know What They Know? How Explainable AI Can Help Us Trust What Machines Learn

Abstract
AI models can make powerful predictions, from diagnosing diseases to forecasting climate extremes, but understanding *why* they make those predictions remains one of the biggest challenges in deploying them responsibly. In this talk, I explore what it really means for an AI system to "know" something.

Using a synthetic benchmark dataset inspired by climate prediction tasks, where the true drivers of the target variable are known, we assess explainable AI (XAI) tools in their ability to recover these drivers under varying levels of noise and data availability – conditions that mirror many real-world scientific and societal applications. Two clear insights emerge: first, explanations become reliable only when models truly learn signal rather than noise; and second, agreement among different explanation methods or models can serve as a practical indicator of whether an AI system has learned something meaningful or is still operating in a state of "epistemic ignorance."

These findings offer guidance for building AI systems that are not only powerful but also aware of their own limits, a key step toward AI that supports science and society with confidence and transparency.

Frequently Asked Questions

1. Is XAI about causality or predictability?

XAI does not directly infer causality. Instead, it helps us understand what patterns in the data a model is using to make predictions, which is about *predictability*, not causal inference.

However, if we use appropriate proxies (such as the ones developed in this work) that measure whether models rely on physically meaningful signals rather than spurious correlations, then XAI becomes a more rigorous tool for generating causal hypotheses. These proxies do not prove causality, but they tell us whether a model is learning something physically real.

2. What does it mean for a model to "perform well for the right reasons"?

Models can sometimes find "loopholes" that allow them to score highly without learning true underlying processes (much like a student getting high marks by guessing long answers on a multiple-choice exam). We want to avoid models that achieve high R^2 through chance correlations or shortcuts.

XAI helps confirm that the model is using relevant, physically meaningful drivers rather than artifacts or spurious correlations. That's what it means to perform well for the *right* reasons.

3. Are XAI heatmaps only for raster data? What about other applications like LLMs?

Heatmaps are most common for rasterized or grid-based data (e.g., images, spatial climate fields). But XAI is used in many other domains as well:

- LLMs use attention maps, token-level attribution, integrated gradients on embeddings, etc.
- Time-series models use temporal attribution.
- Graph neural networks use graph-specific explainers.

So XAI is widely applicable, even if heatmaps themselves are specific to some data types.

4. What's the difference between sensitivity and attribution?

They answer fundamentally different questions:

- **Attribution**: *How much did each feature contribute to the final prediction?*
- **Sensitivity**: *How sensitive is the prediction to small changes in each feature?*

Example:
For $y=\sin(x_1)+\cos(x_2)$ at point (0,0):

- Attribution:
  - $\sin(0)=0 \rightarrow x_1$ contributes 0
  - $\cos(0)=1 \rightarrow x_2$ contributes 1
- Sensitivity (derivatives):
  - $dy/dx_1=\cos(x_1)=1 \rightarrow$ most sensitive to x_1
  - $dy/dx_2=-\sin(x_2)=0$

So, at the point (0,0), the output is attributed entirely to x_2, but is most sensitive to x_1.
These are equally valuable but distinct concepts. Developers of these methods *do* understand this; the confusion usually arises in applied fields (like climate science), where users mix the two and are surprised when results differ.

5. Why do different XAI methods give different explanations?

Because each method:

- makes different simplifying assumptions
- approximates the model in a different way
- captures a different mathematical notion of "importance"

There is no universally optimal XAI method. Their effectiveness varies by task, model architecture, data type, and explanation goal.

6. Should we always use multiple XAI methods?

Based on current evidence: yes.

Different methods highlight different aspects of model reasoning. Using an ensemble of XAI methods (and even an ensemble of AI models) lets you examine the consensus. Structured consensus strongly indicates that the model has learned something real and physically consistent.

7. What is "noise" in data?

Noise can be:

- measurement error
- unresolved physical processes
- variability at scales not captured by the data
- stochastic components of the system

Noise means that the target's variance is not fully predictable. In real-world applications, we never know *how much* variance is ultimately learnable, hence the need for the proxies introduced in this research.

Example:
Winter precipitation forecasting in Southern California is inherently limited. Wintertime small-

scale processes can shift the entire season from wet to dry unexpectedly, meaning the seasonal signal does not encode 100% of the variance. A "perfect" seasonal model would learn all the variance that **is** learnable from seasonal signals, not all the variance that exists.

8. Why does high predictive performance (R^2) not guarantee high XAI fidelity?

Because R^2 measures *outcome correctness*, not *reason correctness*.

A model can achieve high predictive accuracy using spurious correlations or shortcuts. But if it fails to identify the true drivers of the system (low XAI fidelity), then:

- it has not learned meaningful science, and
- it will fail when the environment changes.

Thus, predictive performance is not a reliable proxy for epistemic correctness.

Also, as we mentioned earlier, what high R^2 is depends on how much of the variance is inherently learnable in the system. If for example, for a certain task, only 30% of the variance is learnable, then a model with R^2 on the order of 28% is a great model. In contrast, a model that can capture 60% of the variability in the target out of 85% that happens to be learnable, is not a great model, as there is still a lot of unexplained variance to be learned. Again, in real applications, we never know *how much* variance is ultimately learnable, hence the need for the proxies introduced in this research.

Scientists should aim for models that:

1. achieve the highest possible performance given the data's learnability, and
2. recover the true mechanisms governing the system.

9. Why does XAI consensus matter, and what does it say about meta-cognition?

When models have learned useful information, they show structure and reliability in their explanations. Specifically, we observe that:

- When XAI consensus (across different methods/models) is high, explanations strongly align with ground truth.
- When consensus is low, explanations deviate from ground truth.

This means that AI systems can express confidence in what they know.
When they agree strongly, they tend to be correct; when they do not, they tend to be wrong.

This is a form of meta-cognition: Knowing when you know something, and knowing when you do not.

10. What does "epistemically knowledgeable" mean for AI models?

In classical epistemology, knowledge involves:

- truth
- justification
- reliability

Here, we apply the same idea to AI:

If a model recovers the *true physical reasons* for the target variable's behavior, and its explanations consistently reflect this across ensembles and methods, then it is:

- correct in its predictions
- justified in how it arrives at them
- reliable across settings

Thus, we call such a model epistemically knowledgeable, it "knows" why its predictions are correct.