A black box with a light coming out of it

AI-generated content may be incorrect.

Explainable AI: Making the Black Box Blink

Artificial Intelligence is showing up everywhere these days. It helps doctors diagnose diseases, powers the ads we see online, and even decides if someone gets a loan.

But here’s the catch. When we ask these systems, “Why did you make that choice?” most of them don’t have an answer. Or at least not one we can understand.

That’s where Explainable AI (or XAI) comes in. It’s a way to get AI systems to show their work, to explain how they came up with a decision.

Let’s break it down.

# What Is Explainable AI?

Imagine if your GPS rerouted you through a swamp. When you asked why, it replied, "I’m not at liberty to say." That’s how many AI systems behave.

• Explainability is about making the internal logic of a model clear. It helps us understand how it reached a decision. What information did you use? What parts mattered most? Why this outcome and not another?

• Interpretability focuses on why the outcome matters in context.

Think of it this way: explainability is the process, interpretability is the meaning. Together, they make AI decisions more transparent and digestible.

# What’s a Black Box in AI?

In AI, a “black box” system is one where you can see the input and the output, but what happens in between is a mystery. You don’t know how the model made its decision, even if it works most of the time. This is especially common in complex neural networks, which operate with layers of hidden logic that even their creators often can’t explain. Explainable AI is what cracks open that box and shines a light inside.

# Why Does XAI Matter?

If your AI makes the wrong call, like denying a mortgage or misdiagnosing a tumor, you need to know why. So do regulators, users, auditors, and the people affected.

**Explainable AI helps in several critical ways:**

• It builds trust. We’re more likely to accept AI decisions when we understand them.

• Supports accountability so organizations can take responsibility.

• Enables debugging by tracing errors back to their root causes.

• Helps satisfy regulations such as the EU AI Act and NIST AI RMF.

• Promotes fairness by exposing biased logic or flawed data.

And it helps avoid headlines like “AI System Discriminates Against Job Seekers, No One Knows Why.”

# How Do You Make an AI Explain Itself?

There are two kinds of explainability. Some models are designed to be understandable from the beginning. These are called intrinsically interpretable models, like decision trees, linear regression, or simple scoring systems.

Others, like deep neural networks, need to be explained after the fact using post hoc methods such as LIME (Local Interpretable Model-agnostic Explanations) which explains individual decisions using simpler models, and SHAP (SHapley Additive exPlanations) which shows how much each feature contributed to the outcome, based on game theory, or counterfactuals. In simple terms, some systems are built to show their work, while others need help unpacking it later.

Deep learning models don’t exactly love small talk, but several techniques can help extract meaningful insights:

• Surrogate models create a simpler replica of the original model, often a decision tree, for easier interpretation.

• Counterfactual prompts ask what would have changed the result to illuminate cause-effect logic.

• Gradient-based methods show which input features had the most influence.

• Occlusion analysis hides or removes inputs to see how predictions shift.

• Word clouds and vector visualizations are especially useful in natural language models.

Popular tools like LIME and SHAP are widely used. But beware. They sometimes offer explanations that are technically correct but contextually misleading.

# The Tangled Web of Bias, Hallucinations, and the Limits of XAI

This is where things get complicated.

Bias can sneak in through the data itself. For example, if past hiring decisions favored certain groups, an AI might learn to do the same. Some AI systems still hallucinate (check out our article on AI Hallucinations). They confidently produce fake or misleading answers. Others reflect embedded societal biases in training data. And sometimes the explanations themselves are retrofitted to sound plausible.

Depending on the context, hallucination rates in language models can range from 20 to 46 percent (Meta AI, 2023; Stanford HELM, 2023). That means nearly half of what some AI tools generate could be factually incorrect or entirely made up but still delivered with total confidence. This is not just a glitch. It is a core risk.

In short, XAI can expose flaws, but it cannot always fix them. In some systems, the explanation is just a rationalization. It sounds good, but it doesn’t reveal the true mechanics.

A collage of a person's face and a person's face

AI-generated content may be incorrect.

# Cautionary Tales from the Real World

In June 2023, two New York lawyers were fined $5,000 after submitting a legal brief that included six fictitious case citations generated by ChatGPT. The lawyers admitted using ChatGPT for legal research, unaware that the AI could fabricate cases. The judge criticized them for acting in bad faith and for abandoning their responsibilities.

Read the full article here: <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/>

There is also a case where some cities used AI to predict where crimes might happen. But they couldn’t explain why certain neighborhoods were flagged. There were critiques about the use of data-driven crime prediction tools, arguing that such technologies perpetuate systemic racism and target marginalized communities. The article asserts that predictive policing systems often rely on historically biased data, creating feedback loops that disproportionately impact racialized and low-income groups without demonstrable benefits in reducing crime. ​

Read the full article here: <https://patentpc.com/blog/smart-policing-surveillance-deployment-and-public-safety-stats>

These weren’t bugs. They were failures of governance and explainability.

# Who’s Pushing for XAI?

**Several major frameworks emphasize explainability:**

• NIST AI RMF highlights it as one of the seven pillars of trustworthy AI.

• NIST AI 600-1 emphasizes explainability for generative models to counter hallucinations.

• OECD AI Principles stress transparency and algorithmic accountability.

• GAO AI Framework includes explainability as a requirement for auditing and monitoring.

• EU AI Act requires explainability for high-risk AI applications in finance, justice, and healthcare.

If your AI is making consequential decisions, someone will ask how and why.

# The Honest Truth: It’s Still Hard

Some AI models are just naturally difficult to understand. Often, the more accurate a model is, the harder it becomes to explain. What makes it even more challenging is that different people need different types of explanations. A data scientist might want a breakdown of feature weights, while a doctor or policymaker needs a clear, simple summary. Context plays a big role. A technical explanation may satisfy an engineer but will not mean much to someone whose health or future is on the line.

Explainable AI is not something you turn on like a light switch. It is a gradual process, and it is one that every responsible developer, auditor, and business leader needs to commit to.

# Final Thought: If Your AI Can’t Explain Itself, Can You?

Explainability is not just a technical feature. It is a matter of trust and accountability. When AI systems are involved in decisions about your healthcare, your finances, or your access to opportunities, the logic behind those choices should not be a mystery.

Understanding how an AI came to its conclusion is not a luxury. It is a basic requirement for fairness, oversight, and informed consent. If a system can influence your life, it should at least be able to show its reasoning.

So, the next time someone tells you that AI is just a black box, a fair question to ask is: why is that acceptable?

A person holding a flashlight

AI-generated content may be incorrect.