ESSAY 10

BETTER JUDGMENT IN THE AGE OF AI

An Explanation Is Not a Justification

"Explainability lets you inspect a recommendation. It doesn't validate it."

Alen Mayer

alenmayer.com

An Explanation Is Not a Justification

By Alen Mayer

Explainability has become one of the most trusted words in responsible AI.

A system that can explain what influenced its output feels safer than one that can't, and in an important sense, it is. A black box that simply produces a decision gives a reviewer little to interrogate. A system that shows which factors contributed to its recommendation gives the reviewer something concrete to examine.

What it doesn't automatically give you is an answer to the question that actually matters once you're looking at it: does this conclusion deserve to be acted on?

THE MOST DETAILED REJECTION YOU'LL EVER READ

A bank's AI credit model recommends rejecting a two million euro business loan. The explanation is unusually detailed. A credit officer can see which factors most influenced the recommendation and their relative contribution to the score: cash flow volatility, sector risk, leverage, customer concentration, and a handful of smaller factors.

And the actual decision in front of that credit officer remains. Should the loan be rejected? The explanation answers a different question, why did the model conclude this, with genuine precision. It says nothing about whether that conclusion is the right one to act on.

TWO QUESTIONS THAT LOOK LIKE ONE

Why did the model decide this and should we accept what it decided feel like the same question, because the first one usually seems like it's doing the work of the second. Once you can see what influenced the recommendation, it's natural to feel like you've evaluated the decision. You've looked right at it, after all, and nothing about it seems arbitrary or hidden.

Understanding the first can help enormously with the second. It still doesn't settle it. A credit officer can read the breakdown, confirm it's an accurate representation of what the model weighed, and still have no idea whether cash flow volatility should carry that much weight for this specific business, at this specific moment, or whether these factors and their relative weights are the right basis for this decision at all.

THE EXPLANATION AND THE JUDGMENT ARE DIFFERENT JOBS

Explainability does real work, and it's not nothing: it turns a decision into something that can be inspected, challenged, and audited instead of simply trusted or not. That's a genuine improvement over a system nobody can question. It replaces trust us with here's what most influenced this recommendation, which is the necessary first step before anyone can meaningfully disagree.

But inspection and judgment aren't the same skill, and a transparent model doesn't collapse them into one. Once you can see what drove the recommendation, the harder question opens up rather than closes: does that weighting deserve the influence it was given, in this case?

> Tracing a model's logic and judging whether that logic should govern this particular case are different acts.

The explanation makes those questions possible to ask precisely. It hands you the components instead of a black box. It doesn't answer them.

THE TRAP IS FEELING FINISHED

Here's where it gets genuinely dangerous, not because explainability fails, but because it works so well at the thing it does do. A detailed, accurate explanation feels like due diligence has already happened. The credit officer has seen what influenced the score, confirmed it's coherent, checked that the components add up, and that whole process feels thorough, because it was thorough, at the layer it operated on.

What it can quietly replace is the separate question of whether this decision, in this specific case, is the right one, not because anyone decided that question didn't matter, but because a good explanation is so satisfying to read that it can feel like the harder question already got answered along the way. It didn't. It got described.

WHAT ACTUALLY CLOSES THE GAP

None of this is an argument against explainability. Good explanations can make AI supported decisions easier to inspect, question and audit. The fix isn't less transparency. It's not stopping at transparency.

Before treating an explanation as a justification, the useful question is narrow and specific: does understanding why the model reached this conclusion give me enough reason to believe the conclusion deserves action? Sometimes the answer is genuinely yes. The explanation surfaces something the reviewer wouldn't have caught unassisted, and seeing it is enough to agree the conclusion holds. But sometimes the

honest answer is no. The explanation is accurate and complete and still doesn't settle whether this case is the kind the model's weighting was built to handle well.

Knowing which situation you're in is the judgment call. The explanation can inform that judgment. It cannot, by itself, establish that the recommendation deserves action.

Explainability lets you inspect a recommendation. It doesn't validate it.

TAKE THIS FURTHER

Build better judgment into how your team works with AI.

I help leaders and organizations strengthen the human capabilities behind better judgment as AI becomes part of everyday work and decision-making.

JUDGMENT LABS

A 90-minute session working through realistic AI-influenced business decisions.

WORKSHOPS

Practical training on how teams question, interpret, weigh and decide with AI.

KEYNOTES

Reframing what human judgment requires as AI becomes more capable.

Alen Mayer

alenmayer.com