ESSAY 09

BETTER JUDGMENT IN THE AGE OF AI

# Removing a Variable Is Not the Same as Removing Its Influence

*"Removing a variable removes the variable. It doesn't guarantee it removes what it represented."*

**Alen Mayer**

alenmayer.com

# Removing a Variable Is Not the Same as Removing Its Influence

By Alen Mayer

**Most responsible AI practice includes a seemingly straightforward safeguard: identify information a system should not use, and exclude it.**

Protected characteristics such as race or gender are the obvious examples. Other variables may also be excluded because they risk carrying information the organization does not want influencing the decision.

It's a real safeguard. But removing a field doesn't necessarily remove what that field represented.

## THE RULE THAT LOOKS COMPLETE

The logic behind exclusion is simple enough to feel complete: a system can't discriminate on a characteristic it was never given. Remove the input, remove the risk. Excluding protected or inappropriate inputs is an important safeguard in consequential model design, and it genuinely closes off the most direct form of the problem.

What it doesn't do is guarantee that the pattern the exclusion was meant to prevent doesn't show up anyway, carried in through inputs nobody thought to flag, because each one has a real, independently defensible reason to be in the model.

## CONSIDER THREE VERSIONS OF THE SAME PROBLEM

A regional bank's credit-scoring model excludes race, gender, and zip code by design. It still uses income, employment history, years at a current address, and the selectivity of the applicant's university, each variable with a genuine, independently defensible link to default risk. The bank's fairness audit controls for the model's other major risk factors and still finds a substantial approval gap tied to demographic group. Most of it traces to two variables: university selectivity and address tenure. The model never saw race. It produced a similar pattern anyway.

A hiring algorithm excludes gender entirely. It still weighs continuous years of employment history favorably, and treats gaps in that history as a mild negative signal, a reasonable proxy for consistent experience, on its face. In this scenario, that

scoring also ends up disadvantaging a disproportionate share of candidates whose employment gaps came from parental leave, without anyone designing it to.

An insurance underwriting model excludes race and zip code. It uses distance to the nearest fire station and building age, both genuinely predictive of claims risk. An outcome audit later finds that the resulting pricing pattern is strongly associated with neighborhood demographics. Neither variable was chosen for that reason.

None of these systems cheated. None was explicitly given the excluded characteristic. Each used variables with an independently defensible reason to be there. And each still produced an outcome pattern associated with something the designers had deliberately excluded.

## THIS ISN'T ABOUT ASSUMING BAD FAITH

The uncomfortable finding in each case isn't that the model cheated. It's that the constraint everyone agreed to follow, don't use the protected characteristic, got followed completely, and the outcome it was meant to prevent showed up regardless. That's a harder problem than dishonesty, because there's no dishonesty to find. Nobody violated the rule. The rule just turned out to be narrower than the problem.

> The rule just turned out to be narrower than the problem.

A system can follow every constraint built into it and still reproduce a pattern those constraints were intended to reduce, not because anyone worked around the rule, but because the rule was written about which inputs to exclude, and the outcome runs through relationships between the inputs that remained, none of which broke any rule on its own.

## ABSENT FROM THE INPUTS IS NOT ABSENT FROM THE OUTCOME

The real mistake sitting underneath all three examples is a specific, easy one to make: mistaking absence from the input list for absence from the decision process. You can remove race. You can remove gender. You can remove the zip code field entirely. What you can't do by removing a field is guarantee that nothing left behind still carries the information that field represented, because real-world variables are correlated with each other for genuine historical and structural reasons that have nothing to do with any model. Removing the field doesn't remove those relationships.

That's the trap specifically. Exclusion feels like it settles the question, because it's a concrete, checkable action: the variable is gone, verifiably, and that verification is satisfying in a way that "audit the outcomes" isn't. But checking what went into the model and checking what the model produces are two different exercises, and only one of them tells you whether the exclusion actually worked.

## WHAT ACTUALLY HELPS

This doesn't mean every correlation is secretly a proxy, or that models should stop using legitimate risk factors because they happen to correlate with something protected. University selectivity may earn a meaningful share of its predictive power honestly. Employment continuity may too. Treating every correlated variable as automatically tainted would gut models of real signal for reasons that don't hold up: correlation on its own proves nothing either way.

What actually helps is checking, deliberately, rather than assuming the exclusion settled it: does the outcome, in aggregate, reproduce the pattern the exclusion was meant to prevent? How much predictive value does the correlated variable retain once its relationship with the protected characteristic is accounted for: is it contributing useful independent signal, or is some of its apparent value coming from the pattern you're trying not to reproduce? And if the answer isn't clear, uncertainty doesn't justify either conclusion by default. It just means more work is required before the organization can claim to understand what the variable is doing.

Checking the inputs is not enough. Judgment also has to examine what those inputs, combined, actually produce.

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**