

BEYOND MODEL ACCURACY

Accuracy Was Never the Bottleneck: The Human Side of AI Adoption Everyone Skips

Why the models we validate so carefully so often change nothing — a view from the enterprise delivery floor.

Giovanni Leonardi

August 2022



Accuracy Was Never the Bottleneck: The Human Side of AI Adoption Everyone Skips

Transformation · No. 283

Originally written August 2022 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

"A model does not fail when it is wrong; it fails when it is right and no one acts on it."

1. The launch that changes nothing

There is a particular kind of silence that follows a successful model launch. The build is done. The validation pack is signed off. The accuracy numbers are better than anyone dared hope at kick-off, and someone has put them on a slide with a green tick. The steering group applauds. The data science lead is thanked by name. And then, over the weeks that follow, the thing that was supposed to change the business quietly does not.

The dashboards still get built. The model still scores every case overnight. But the underwriter still backs her own judgement, the planner still runs the spreadsheet he has trusted for nine years, and the branch manager still asks the analyst for "the real number" before he acts. Six months on, someone finally pulls the usage logs, and the awkward truth surfaces: the model that was validated at the top of its class is being consulted on barely one decision in ten.

I have watched this sequence play out enough times to stop treating it as bad luck. It is not a data problem, and it is almost never a modelling problem. It is the predictable result of pouring months of effort into making a model correct and almost none into making it used. We are, as a profession, fluent in the first task and strikingly amateur at the second.

2. Accuracy is a solved problem; adoption is not

The uncomfortable thing to say out loud in 2022 is that model accuracy has largely stopped being the constraint. The tooling has matured to the point where a competent team, given decent data, will produce a model that beats the incumbent process on almost any metric you care to name. Feature stores, managed training pipelines, the steady professionalisation of what people are starting to call MLOps — all of it has made the building of a defensible model a repeatable exercise rather than a heroic one.

Adoption has not followed the same curve. If anything the gap has widened, because every gain in how quickly we can ship a model is a gain in how quickly we can ship a model that no one uses. The bottleneck has moved. It now sits precisely where we are least equipped to work: in the heads of the people who were supposed to change what

they do.

A model does not fail when it is wrong; it fails when it is right and no one acts on it. Those two failures look nothing alike on a project plan. The first shows up in test metrics and gets fixed before go-live. The second shows up months later in a usage report nobody commissioned, long after the delivery team has rolled off and the budget has closed. One is designed for; the other is left to chance.

The question that decides whether an AI initiative was worth doing is not "Is the model good?" but "Did anyone change what they do because of it?" We instrument the first question obsessively and the second one almost never.

3. Why the human side gets skipped

If the human side of adoption matters so much, why is it the first thing to fall off the plan? Not through ignorance — most experienced sponsors will nod along if you tell them people matter. It gets skipped because of three structural forces that are stronger than good intentions.

The first is who holds the budget and how it is drawn down. AI initiatives are funded as technology builds. The line items are data engineering, platform, model development, integration. "Getting three hundred underwriters to trust and act on a score" is not a line item anybody knows how to cost, so it becomes an afterthought funded from whatever is left — which, by the time the build has run its course, is nothing.

The second is a skills asymmetry. The people who are brilliant at building the model are rarely the people who are good at changing an operating routine, and we staff these initiatives almost entirely with the former. A team that can reason about cross-validation folds has no particular reason to be able to reason about why a claims handler distrusts a number she cannot interrogate. Neither should we expect it to. But we act surprised when a room full of model builders does not, unprompted, do the work of behavioural change.

The third, and the deepest, is a category error about what kind of problem adoption is. We treat "roll-out" as the tail end of a delivery — training slides, a launch email, a floor-walker for the first week — as though the challenge were informational, a matter of telling people the tool exists. It is not informational. It is a matter of trust, of professional identity, and of whether acting on the model makes a practitioner's day better or worse. Those are not solved by an email.

- We fund the model, not the adoption of it, and then measure the model rather than the adoption.
- We staff for building, not for behaviour change, and then wonder at the behaviour.
- We treat resistance as ignorance to be corrected, when it is usually information to be understood.

4. What actually earns a practitioner's trust

The word that gets used, when a model is ignored, is resistance — and it is almost always the wrong word. When you actually sit with the underwriter who overrides the score six times out of ten and ask her why, you rarely find stubbornness. You find reasons, and the reasons are usually good ones.

She has been accountable for these decisions for a decade; the model has been accountable for none of them. When it is wrong, she carries it, not the model. She cannot see why it produced a given score, so she cannot tell the difference between a case where it is confidently right and one where it is confidently wrong — and in her world those are not the same risk at all. And acting on it, in the workflow as built, takes her more clicks and more time than ignoring it. Every one of those is a rational objection, and none of them is addressed by telling her the model is 82% accurate.

What earns trust is the opposite of a launch email. It is the slow accumulation of evidence that the thing is worth acting on, delivered in the terms the practitioner actually cares about.

What we measure about a model	What makes someone use it
Accuracy, AUC, precision/recall	Whether it is right on the cases that matter to me
Overall error rate	Whether I can tell when to trust it and when not to
Coverage of the population	Whether acting on it is easier than ignoring it
Time-to-deploy	Whether I carry the blame when it is wrong
Model documentation	Whether someone I respect already relies on it

None of the right-hand column is a modelling property. All of it is a property of how the model meets the person — its explanation, its fit to the workflow, the recourse when it errs, the accountability around it, and the social proof of a respected peer who already leans on it. That is the actual product. The model is only its engine.

"Resistance is not the opposite of adoption. It is the most honest market research you will ever get, delivered free, by the only people whose behaviour the whole initiative depends on."

5. "The technology will make this go away"

The strongest objection to everything above is worth stating in its full force, because it is not a foolish one. It goes like this: adoption is a friction problem, and friction is

temporary. The models are getting better and the interfaces are getting smoother. The newest generation of large models can already draft text and write usable code; before long these systems will be good enough, and embedded smoothly enough, that using them will be effortless and trusting them will be automatic. The human problems you are describing are the growing pains of an immature technology, and technology outgrows its growing pains.

There is real truth in it. Better models and better interfaces do lower the cost of adoption, and some of what looks like resistance today genuinely is friction that will be engineered away. It would be foolish to deny it.

But it mistakes the nature of the problem. Trust is not friction. Making a tool faster and more fluent does not, on its own, make a professional willing to stake her name on its output — and in many cases it makes her more wary, because a system that is confident and fluent and wrong is more dangerous than one that is obviously clumsy. The more capable these systems become, the more the decisive question shifts away from can it produce an answer toward should this person act on this answer, here, accountable for this outcome — and that question is irreducibly human. Capability raises the stakes of the trust problem; it does not dissolve it. Betting your transformation on the technology outgrowing the need for change management is, in my experience, the single most expensive assumption an AI programme can make.

6. A model that finally got used

Let me make this concrete, because the argument earns nothing if it stays in the abstract.

Consider a claims operation — the shape of this recurs across insurers, and the specifics here are composited from the pattern rather than drawn from any one of them. A triage model was built to flag which incoming claims could be fast-tracked and which needed a full assessor's review. It was good: on back-testing it agreed with expert assessors on the clear-cut cases and freed an estimated fifth of assessor time. It went live to a room of forty assessors. Ninety days later, fast-track usage sat at eleven percent, against a business case that had assumed sixty. The model was not wrong. It was unused.

The instinct in the room was to blame culture and order more training. What actually turned it around was a change of subject — from the model to the people the model was meant to serve. Three things were done, none of them data science.

1. The score was made legible. Instead of a bare number, each flagged claim carried the two or three factors that had driven it, in the assessors' own vocabulary. Not a full explanation — just enough for an experienced hand to sanity-check the machine against her own instinct in five seconds.
2. The accountability was moved off the individual. A fast-track decision the model recommended and the assessor accepted was owned by the process, signed off in policy by the operations director. The assessor was no longer personally carrying a machine's call.
3. The workflow was inverted so the model's path was the path of least resistance. Accepting the fast-track was one click; overriding it took a short note. The friction was deliberately moved onto the exception, not the rule.

Within a quarter, fast-track usage rose from eleven percent to just under seventy, and the assessors — the same "resistant" assessors — were the ones defending the model when a regional manager questioned it. Nothing about the model had changed. Everything about how it met the people had.

The point of the story is not the three specific moves; another operation would need three different ones. The point is that every one of them was invisible to a plan that defined the work as building a model and the roll-out as telling people about it.

7. The work nobody budgets for

So here is the position, stated plainly. The scarce capability in AI adoption is no longer the ability to build a good model. It is the ability to change what a competent, accountable professional does on a Tuesday afternoon — and that is a discipline of its own, as demanding as the modelling and far less staffed. Treating it as an afterthought is not a small planning error. It is a decision, made implicitly, to spend the whole budget building an engine and none of it connecting the engine to the wheels.

If I were to argue for one change in how these initiatives are set up, it would be this: fund and staff the adoption of a model as seriously as its construction, from the first day and not the last. Put someone accountable for use, not just delivery, in the room from kick-off. Instrument the question "did anyone change what they do?" with the same rigour we instrument accuracy. And treat the first wave of resistance not as a communications failure to be managed away, but as the clearest signal you will get about what the model must earn before it deserves to be trusted.

The organisations that get value from AI over the next few years will not, on the whole, be the ones with the best models. On current trends good models will be widely available. They will be the ones that took the human side seriously enough to fund it — the ones that understood, before they spent the money, that a model changes nothing until a person decides to act on it, and that getting them to decide is the actual work.