

BEYOND THE HOLDOUT SET

When Models Met Governance: The Accountability Gap in Production Machine Learning

A reflection on the collision between data science and the disciplines of risk, audit, and regulation — and what it exposed about how far the data-driven enterprise had actually come.

Giovanni Leonardi

October 2019



Contents

1. Executive Summary	3
2. The Question That Stops the Room	4
3. Two Meanings of the Word "Done"	4
4. Why the Gap Would Not Close	5
5. The Case for Leaving Well Alone	6
6. Governing the Decision, Not the Experiment	7
7. The Seam in the Transformation	8

When Models Met Governance: The Accountability Gap in Production Machine Learning

Transformation · No. 234

Originally written October 2019 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

“A model was the rare artefact in the enterprise that could decay without anyone touching it, because the world it had learned kept moving underneath it.”

1. Executive Summary

Somewhere between 2016 and 2019, machine-learning models crossed a line that few organisations had thought to mark. They moved out of the analyst's notebook and into production, where they began making thousands of consequential decisions a day — which application to decline, which transaction to flag, which customer to keep. What met them there was not more engineering. It was governance: the accumulated apparatus of risk management, internal audit, change control, and, from May 2018, data-protection law with real teeth. The two did not fit together, and the friction between them became one of the defining and least understood experiences of the period.

This essay argues that the friction was not a misunderstanding to be smoothed away by better communication. It was a structural collision between two incompatible definitions of when a piece of work is finished. Data science declared a model done when it was accurate on data it had not seen. Governance declared it done when it was documented, attributable, reproducible, monitored, and reversible. Neither definition was foolish. The distance between them was built into the organisation chart, into the tooling, and into the mathematics itself.

We trace why the gap would not close on its own: models were built outside the disciplines that governed everything else the enterprise did; a trained model turned out to be a fragile compound of code, data, and configuration that most teams versioned only in part; and a model was the rare artefact that could degrade in production without anyone touching it. We take seriously the strongest case against governance — that disciplining machine learning too early destroys the very iteration speed that makes it valuable — and argue that the resolution lies in governing the decision rather than the experiment. Finally, we read the collision as a diagnostic. The organisations that had proclaimed themselves data-driven had, for the most part, re-tooled and re-scripted without rebuilding their structures of accountability. The models had run ahead of the institution. What was only beginning to close that distance is described here from inside the autumn of 2019 — as a set of nascent practices whose adequacy remained an open

question.

2. The Question That Stops the Room

Every practitioner who lived through this period has a version of the same scene. A model has been in production for the better part of a year. It is, by the only measure anyone applied when it shipped, a success: its accuracy on the validation set was the best the team had produced, and it has been quietly scoring cases ever since — tens of thousands of them — with no one paying it much attention, because software that works is software you stop watching.

Then something arrives from outside the team. It might be a complaint escalated by the contact centre. It might be a letter from a regulator. It might be a single line in an internal audit's terms of reference: on what basis are these decisions made? And a meeting is called, and someone asks a question that sounds simple and turns out to be unanswerable. Why was this particular person declined?

The honest answer, in the room, is that the model declined them. But that is not an answer anyone can write down and stand behind. So the team goes looking for the reason, and discovers that reconstructing a single decision made eleven months ago is close to impossible. The model has been retrained twice since. The training data behind the original version has been overwritten by the newer snapshots. The notebook that produced the first version lives on the laptop of an analyst who has since left. And the feature that weighed most heavily in the decision was a ratio derived from two other engineered features, with no plain-language meaning that could be offered to the person who was refused.

Nothing here involves a bug. The system performed exactly as designed. What the room has discovered is not a defect in the model but a gap in the enterprise — the absence of any machinery for answering, after the fact, a question that the organisation has always been able to answer about every other kind of decision it makes. Underwriters could explain a decline. Credit committees kept minutes. The model kept none.

3. Two Meanings of the Word "Done"

The deepest source of the friction was that the two communities meeting over these models were each using the word done to mean something the other did not recognise.

For the data scientist, a model was finished when it generalised — when it performed well on data held back from training, when the error curves had flattened, when the next increment of effort bought a rounding error's worth of accuracy. This was not laziness; it was the discipline of the field, and a real one. The holdout set was an honest test, and shipping was the point. A model that never left the notebook created no value at all.

For the risk manager, the auditor, the compliance officer — the people we might loosely call the governors — a model was finished when it could be defended. Not when it was accurate, but when there was a named owner accountable for it, a record of how it was built, a way to reproduce any decision it had made, a monitor watching it in life, and a

route to switch it off. Their discipline, too, was real, and older than machine learning by decades. Banks had run model-risk functions since well before anyone spoke of data science; the regulatory expectation that a firm should validate, document, and independently review its models was, by 2019, close to a decade into codified practice in financial services.

Dimension	"Done" for the data scientist	"Done" for the governor
The primary test	Accurate on unseen data	Accountable to a named owner
The unit of value	The model's performance	The decision's defensibility
Attitude to change	Iterate fast; retrain often	Control change; preserve the record
The time horizon that matters	The validation set, today	The audit, eighteen months from now
The failure most feared	A worse score	An outcome no one can explain

Set out this way, the collision looks less like a clash of competence and more like two professions optimising honestly for different things. The data scientist feared shipping something that did not work. The governor feared being unable, one day, to account for something that had. Both fears were legitimate. The trouble was that each side experienced the other's discipline as obstruction — the governors saw cowboys shipping unexplainable decisions into live operations, and the data scientists saw a bureaucracy demanding paperwork about experiments that would be thrown away next week.

4. Why the Gap Would Not Close

Had this been merely a cultural misunderstanding, it would have yielded to the usual remedies — a shared workshop, a translation layer, a diplomat who spoke both languages. It did not yield, because the gap was held open by structural forces that no amount of goodwill could dissolve.

- The models were built in the wrong place on the org chart. Machine learning grew up inside analytics teams, innovation labs, and digital functions — precisely the parts of the organisation that had been set up to move fast and stand outside the change-control and model-risk disciplines that governed the core. The very structure that let these teams produce models quickly was the structure that kept them beyond the reach of the governance that would later come looking. The gap was drawn into the organisation chart before the first model shipped.
- A trained model is not one artefact but four, and most teams kept only one. A model in production is the product of code, training data, a configuration of hyperparameters, and the specific run that combined them. To reproduce a decision you need all four, versioned together. Engineering culture had taught everyone to version code; almost no one, in the early days, versioned the training snapshot and the run that went with it. So the moment a model was retrained, its earlier self became irrecoverable — and with it, the ability to explain anything the earlier self had

done.

- A model decays without being touched. This was the strangest property, and the one governance was least prepared for. Every other artefact the enterprise controlled stayed put once deployed; a payroll rule did not quietly become less true over the summer. A model did. It had learned the world as the world was during training, and the world kept moving. A model was the rare artefact in the enterprise that could decay without anyone touching it, because the world it had learned kept moving underneath it. Consider a fraud model that scored strongly on its validation data in January and was, by September, catching little more than half of what it had caught at launch — not because a line of code had changed, but because the fraud had. Governance was built to control change. Here was an artefact that changed the most precisely when nobody changed it.
- The most accurate models were often the least explainable. And now the law was asking for explanations. The data-protection regime that took effect in 2018 gave individuals rights over decisions made about them by automated means, and turned why from an engineering nicety into a legal question. The techniques then emerging to prise open a model's reasoning — local approximations, feature-attribution methods that assigned each input a share of the outcome — were genuine advances, and none of them delivered the plain, causal, human-legible reason that an adverse-action notice had always implied. There was a real trade-off between how well a model performed and how readily it could be accounted for, and the field had spent years optimising hard for one side of it.

The organisations did not lack diligence. They had built a new kind of author of decisions — a statistical process trained on data — and had simply never asked what it would mean to be accountable for what that author decided.

5. The Case for Leaving Well Alone

It would be easy, and wrong, to tell this story as one of reckless engineers finally brought to heel by prudent governance. The strongest voices on the other side were not reckless, and their argument deserves to be met at its best.

That argument runs as follows. Machine learning creates value through iteration. A model is not a bridge, designed once and built to specification; it is an experiment that improves by being tried, measured, and revised, often many times a week. Subject that loop to the full weight of traditional model governance — sign-offs, documentation gates, independent validation before any change — and you do not make the model safer; you make it stop. The cost is not theoretical. In a domain where a competitor willing to tolerate a little more model risk can iterate ten times for each of your one, heavyweight governance is not caution but slow-motion surrender. And much of what the governors demanded was, on this view, a category error: you cannot keep minutes for an experiment, cannot write an audit trail for a hypothesis you intend to discard on Friday. Impose the apparatus of permanence on something whose whole value is its impermanence, and you have misunderstood the thing you are governing.

This is a serious case, and the weakest response to it is the one most often given at the time — that safety simply matters more than speed, so the process must be borne. That answer loses the argument, because it refuses to weigh the real cost of shipping nothing.

6. Governing the Decision, Not the Experiment

The way through, where organisations found one, lay in noticing that the objection and the requirement were aimed at two different objects that everyone had been treating as one.

There is the experiment: the notebook, the feature that might not survive the week, the twelfth retraining of an idea still looking for signal. And there is the decision: the moment a model's output stops being a number on a data scientist's screen and starts changing what happens to a real person — an application refused, a payment held, a claim routed to investigation. These are not the same object, and they do not warrant the same treatment.

The mistake almost everyone made in the first years was to govern both the same way — which in practice meant governing neither, until an incident, and then governing everything at once in a panic. The experiment got no oversight because oversight would have killed it; and because the experiment got none, the decision inherited none, and sailed into production ungoverned. The reconciliation that began to take shape was to draw the line not around the technology but around the consequence.

1. Let the experiment stay fast and lightly held — versioned for the team's own sanity, but free of ceremony, because nothing it does yet touches anyone.
2. Treat the crossing into production, where a model begins to affect real people, as the governed event — the point at which a named owner, a reproducible record, a live monitor, and a documented basis for the decision become non-negotiable.
3. Match the weight of governance to the stakes of the decision, not to the sophistication of the method — a model steering a marketing email and a model refusing credit are the same technology and entirely different accountabilities.

“The model had not broken the rules. It had revealed that the enterprise's rules were written for decisions with a human author, and had never contemplated any other kind.”

Stated this way, the governor's instinct and the data scientist's instinct stop being enemies. The demand for reproducibility, monitoring, and a defensible basis is not a demand made of the experiment; it is a demand made of the decision, and it is exactly as reasonable for a model as it always was for a person. What changed was only that the author of the decision was now a trained function rather than an underwriter — and the enterprise had built no chair for it to sit in, no signature for it to sign.

7. The Seam in the Transformation

Step back from the particular quarrels — the unanswerable meeting, the overwritten training data, the drifting fraud model — and the episode reads as a diagnostic of something larger, and less comfortable, about the transformations of the period.

Through these years, a great many organisations declared themselves to be becoming data-driven. They meant it, and they invested accordingly: platforms were bought, data scientists were hired, dashboards multiplied, and the vocabulary of the enterprise filled with models and pipelines and features. Most of this was real. But it operated almost entirely at the level of tooling and rhetoric, and the collision with governance exposed how little of it had reached the level of accountability. Becoming data-driven had been understood as acquiring a capability. It had not been understood as taking on a new kind of responsibility — the responsibility for decisions produced by a process no single person authored and few could fully explain.

That is the seam the collision exposed. The models were, in a precise sense, ahead of the institution that housed them. The technical transformation had genuinely happened; the institutional transformation — the slower, duller work of deciding who answers for a machine's decision, how it is recorded, how it is watched, how it is switched off — had barely begun. And the gap between the two was not a lag that would close by itself with a little more time. It was the visible edge of the difference between changing what an organisation does and changing what it is accountable for, which are not the same transformation and are rarely done by the same people.

By the autumn of 2019, the outlines of a reconciliation were becoming visible, though it would be dishonest to call them settled. A discipline was beginning to cohere around the operation of models in production — the monitoring of live performance, the versioning of data alongside code, the practice of documenting a model's intended use and known limitations in a standard form. Some of it was borrowed, sensibly, from the model-risk functions that banking had run for years; some was genuinely new. Whether this emerging practice would harden into real accountability or settle into a fresh layer of theatre — documentation produced to satisfy an audit and read by no one — was, at the time of writing, entirely undecided. What could be said was narrower, and it is where this reflection ends. The organisations that fared best were not those with the most accurate models or the heaviest process. They were the ones that had understood, earlier than the rest, that the moment a model begins to decide is the moment someone must be able to answer for it — and had started building that answer before the room went quiet and the question was asked for them.