

BEYOND EXPLAINABILITY

The Framework Always Arrives Last: Algorithmic Accountability in the Gap Before the Rules

What the long history of accountability catching up to new power tells us about the algorithms now deciding who gets the loan, the job, and the benefit of the doubt

Giovanni Leonardi

January 2019



Contents

1. Executive Summary	3
2. The decision no one could explain	4
3. The long habit of arriving late	4
4. Why this time is harder	5
5. The comfort of a technical answer	6
6. Four things we call one	7
7. A question of temperament	8
8. The longer view, again	8

The Framework Always Arrives Last: Algorithmic Accountability in the Gap Before the Rules

Transformation · No. 277

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

“Accountability has never arrived on time; it arrives after the harm that makes it unavoidable, and the only choice a profession really has is whether to build it a little ahead of that moment or a long way behind it.”

1. Executive Summary

A decade from now we will have frameworks for algorithmic accountability — standards, impact assessments, guidance by the shelf-load, perhaps an inspectorate. We do not have them yet. In the meantime the models are already live: deciding who is offered credit, whose job application is ever seen by a human, whose insurance claim is paid without a second glance and whose is held back for scrutiny. The decisions are in production. The accountability is still on the whiteboard.

This essay takes the long view of that gap, and argues it should not surprise us. Accountability has always arrived after the power it governs, never before it. Double-entry bookkeeping and then audited accounts followed the joint-stock company by generations; product-safety law followed mass manufacture; clinical governance followed the medicines it now restrains. We are living through the same lag with automated decisions — but with three features that make this case harder than its predecessors. The decisions are opaque even to the people who build them. They operate at a scale no earlier accountability regime ever imagined. And they arrive dressed in the borrowed authority of mathematics, which makes them unusually hard to question.

The reflex response — a scramble toward technical explainability, and a hopeful reading of new data-protection law as though it already settled the matter — mistakes a governance problem for an engineering one. Explanation is necessary and nowhere near sufficient. What we loosely call accountability is really four different things bundled into one word: the ability to explain a decision, the ability to contest it, a named party who answers for it, and a route to redress when it turns out to be wrong. A tool that generates a plausible reason for a rejection delivers the first and quietly abandons the other three.

- Accountability is always retrofitted; the only real question is whether it arrives a little before the harm or a long way after it.

- Automated decisions are a harder case than the historical precedents, because of opacity, scale, and the false comfort of apparent objectivity.
- Explainability techniques and today's reading of data-protection law treat the symptom, not the structure — and can manufacture a reassurance that makes matters worse.
- The organisations that will look prudent in hindsight are building the unglamorous machinery now, before anyone compels them to.

We are, as a profession, fluent in building these systems and far less fluent in restraining them. That imbalance — not any shortage of clever technique — is the accountability gap.

2. The decision no one could explain

Picture the meeting. A good candidate has been rejected, and someone senior wants to know why. The recruiter did not reject her; a model did, before any recruiter saw the file. The screening system had been trained on ten years of the organisation's own hiring decisions and set loose to rank incoming applications. It now filters four of every five applicants before a human opens a single file, and its choices correlate respectably with the ones the firm used to make by hand. Asked why this candidate, the team can offer a score and a threshold. Asked what in her application produced the score, they can offer a shrug dressed as a sentence: the model weighs many features; no single one is decisive; the pattern is what matters.

This is the ordinary texture of the problem in practice — not a dystopia, just a Tuesday. The system is not malicious and often not even wrong in aggregate. It is simply unaccountable in the particular, and the particular is where people live. The candidate does not experience an aggregate. She experiences a door that did not open, and a reason no one in the building can give her.

Multiply that meeting across credit, insurance, fraud, benefits, and the shortlisting of almost everything, and you have the defining governance question of the moment — arriving, as such questions do, well ahead of any settled means of answering it.

3. The long habit of arriving late

Step back far enough and the present anxiety loses its novelty. New power has always outrun the machinery built to hold it to account, and that machinery has always been assembled afterwards, usually in the wake of a visible harm.

The joint-stock company let strangers pool capital at a scale no partnership could, and for a long stretch there was almost nothing we would now recognise as financial accountability; audited accounts, disclosure rules, and the whole apparatus of the annual report were retrofitted onto an institution that had already reshaped the economy. Mass manufacture put powered machinery and novel chemistry into daily life decades before the safety regimes that now govern them; the standards followed the accidents. Medicine gave itself the power to intervene profoundly in the body well before the ethics committees, trial protocols, and consent doctrines that now discipline it — most of that

scaffolding is a twentieth-century response to twentieth-century abuses.

The pattern is consistent enough to be treated as a law rather than a coincidence: the capability is adopted because it is useful, the harms surface only at scale and only after adoption, and the accountability is built last, by people cleaning up after a technology that has already moved on. Seen this way, the position we are in with automated decisions is not a strange new failure. It is the normal condition of a powerful tool in the years before its reckoning.

That is oddly consoling and profoundly not. Consoling, because it tells us the framework will come; the absence of one today is not evidence that the problem is unsolvable. Not consoling at all, because the same history tells us what usually summons the framework into being. It is rarely foresight. It is usually a scandal large enough that inaction becomes the greater risk.

The framework always arrives. The only variable a profession actually controls is whether it arrives a little ahead of the harm, built deliberately, or a long way behind it, written in anger.

4. Why this time is harder

If the lag were the whole story, we could simply wait for the familiar cycle to complete. But automated decision-making carries three features that make it a harder case than the joint-stock company or the factory floor, and each one weakens the tools we would normally reach for.

The first is opacity. In earlier cases the mechanism could in principle be inspected: a ledger can be read, a machine dismantled, a protocol audited against what was actually done. A model built by fitting millions of parameters to historical data is different in kind. Its logic is distributed across the whole, resistant to the question why this one in a way a rulebook never is. This is not a temporary limitation to be engineered away by next year; for the most capable methods it is close to constitutive. We have built decision-makers we cannot fully interrogate, and then placed them exactly where interrogation matters most.

The second is scale. A biased loan officer harms the people who reach that officer's desk. A biased model harms everyone the system touches, identically, instantly, and invisibly — and it does so with a consistency we are tempted to mistake for fairness. Uniform error at national scale is not a smaller problem than scattered human prejudice; it is a larger one wearing the costume of rigour.

The third, and most corrosive, is the borrowed authority of mathematics. A number arrives looking objective. It does not look like the judgement it in fact is — a judgement about which data to gather, which outcome to optimise, which historical pattern to treat as ground truth. The past few years' most useful correction has been the growing recognition that a model trained on the record of how we behaved will faithfully reproduce how we behaved, prejudices included, and hand the result back to us stamped with the neutrality of arithmetic.

"The maths did not remove the judgement; it removed the judge."

That is the trap in a sentence. The judgement is still there, dense with human choices, but the human who might once have been asked to defend it has been dissolved into a pipeline. Accountability needs someone to answer. The system's great convenience is that, by construction, no one need.

5. The comfort of a technical answer

Faced with this, the instinct of a technical profession is to reach for a technical remedy, and two are currently on offer.

The first is explainability: a fast-growing body of methods that sit alongside an opaque model and produce, for any given decision, an account of which inputs seemed to push it one way or another. The honest case for these tools is real, and worth stating at its strongest. They can surface a variable that should never have been in the model. They can give a reviewer somewhere to start. They can turn a bare score into something a person can at least argue with. In a field with too little transparency, more is genuinely better, and the researchers building these methods are doing serious work.

But we should be clear about what an explanation of this kind is. It is a plausible story about a decision, generated after the fact — not the reason the decision was made, because for these models there is no single reason of the sort a human account demands. The danger is not that the tool is useless; it is that it is reassuring. A well-formatted explanation lets an organisation feel it has discharged its duty when it has merely narrated its output. Worse, the explanation can be perfectly faithful to the model and still damning — "the application scored lower because of a two-year gap in employment" is an accurate account, and if that gap reflects parental leave it is a confession rather than a defence. Explainability tells you what the model did. It is silent on whether the model should have.

The second technical comfort is legal, and specific to this moment. The new data-protection regime that came into force across Europe last May is widely read as having created a "right to explanation" for automated decisions — a right for the individual to be told the logic involved and to have a human review the outcome. Practitioners speak of it as if it settles the question. It does not. Whether the regulation confers a genuine, enforceable right to a meaningful explanation, or something considerably softer, is a matter of live and unresolved argument among the people who read these texts for a living. The provisions gesture at meaningful information about the logic; they stop well short of prescribing what would make an automated decision truly answerable. Leaning the whole weight of accountability on a contested reading of a young regulation is not governance. It is hoping the lawyers will save us from a problem we would rather not solve ourselves.

Here is the strongest version of the opposing view, and it deserves a hearing. Heavy governance, imposed now, will simply slow us down while more aggressive competitors race ahead; and in any case the technology will mature — better methods, clearer case law, the ethics guidance now circulating in draft — so the problem will ease of its own accord. There is something to this. Premature, box-ticking bureaucracy can smother useful work without protecting a single person, and much of what currently travels under the banner of "AI ethics" is exactly that. But the argument proves less than it claims. Waiting for the technology to mature is precisely the bet the long history warns against, because maturity has never arrived ahead of the reckoning — it has arrived because of it. And the competitive case cuts both ways: the reputational cost of a visible algorithmic harm, the automated decision later shown to have discriminated quietly and at scale, is now large enough that restraint is not obviously the slower path. The real choice is not governance versus speed. It is deliberate accountability now versus imposed accountability later, on someone else's terms and timetable.

6. Four things we call one

The deeper reason the technical answers disappoint is that they address one part of a problem we persist in treating as indivisible. The word accountability does a great deal of quiet work, and it is worth prising apart, because most current effort lands on the first element and mistakes it for the whole.

Element	The question it answers	What it actually requires
Explanation	Why was this decision made?	An account a non-specialist can understand and challenge — not a technical trace
Contestability	How do I push back?	A real, accessible route to have the decision re-examined, with the burden off the person harmed
Responsibility	Who answers for it?	A named human or role that owns the decision and cannot dissolve into "the system"
Redress	What happens when it is wrong?	A means of correction and remedy, and a feedback loop so the same error is not repeated tomorrow

Set out like this, the imbalance is obvious. Explainability tooling and the hopeful legal reading crowd around the first row. The other three — contestability, responsibility, redress — are barely technical at all. They are organisational commitments: someone must own the decision, someone must staff the appeal, someone must pay for the correction and close the loop. None of that is delivered by a cleverer method. All of it costs money and attention and a willingness to be answerable, which is exactly why it is so often deferred in favour of the tool that promises to make the discomfort disappear.

What does it look like to take all four seriously, before anyone forces the issue? Concretely — and this is a composite of what more thoughtful organisations are already feeling their way toward — a model that materially affects people passes an impact assessment as routine as a financial sign-off before it goes anywhere near production: what decision it makes, on whom, what its failure modes are, and who is accountable

when they occur. That assessment has a named owner, not a committee that convenes when convenient. It sets a threshold — that any decision above a defined line of consequence cannot be fully automated without a human who can be asked to account for it. It builds the appeal route before launch, not after the complaints arrive, and it resources that route rather than gesturing at it. And it treats every overturned decision as data, feeding the reversals back so that a model which wrongly penalised the two-year employment gap is corrected — rather than left to reject next year's returning parents in the same silence.

None of this is exotic. It is closer to health-and-safety than to computer science — dull, procedural, and precisely the sort of unglamorous scaffolding that every mature accountability regime eventually turns out to be made of. Its absence today is not a technical gap. It is a choice we are making by default.

7. A question of temperament

Which brings the argument to its uncomfortable centre. The reason accountability lags is not, mostly, that the methods are missing. Impact assessments are not hard to design. Naming an owner is not a research problem. The lag persists because building these systems is thrilling and restraining them is tedious, and organisations reliably staff the thrilling work and starve the tedious.

We are fluent in capability and inarticulate in restraint. The analytics team is measured on models shipped and minutes saved, never on harms averted, because averted harms are invisible and shipped models are not. The incentive gradient runs entirely one way. Everyone can see the efficiency the model delivers this quarter; no one is rewarded for the scandal that did not happen because someone insisted on an appeal route that has not yet been used. Accountability, in this light, is not primarily a technical or even a regulatory challenge. It is a challenge of temperament and of incentive — of whether an organisation can bring itself to value the quiet discipline that produces nothing visible except the absence of disaster.

I have sat through enough of these conversations to recognise the tell. The moment governance is raised, the discussion slides with practised ease from should we to do we have to — from the ethical question to the compliance one, from what we owe the person on the other side of the decision to what the regulator will eventually make us do. That slide is the whole problem in miniature. It is the sound of a profession waiting to be forced, when it could still choose.

8. The longer view, again

None of this will be got right the first time. The frameworks we eventually build for automated decisions will be clumsy — over-broad in places, full of holes in others — and a later generation will look back on our early attempts much as we now look back on the first company law or the first factory inspectorates: as well-meant, necessary, and embarrassingly crude. That is not a reason to wait for a better moment. There is no better moment. There is only the ordinary condition of every powerful technology in the years

before its reckoning, and the single decision a profession actually gets to make within it.

Because the framework is coming. The long history is clear on that, and clear too on how it usually comes: late, and in anger, summoned by a harm large enough that no one can any longer pretend not to have seen it approaching. Accountability has never arrived on time; it arrives after the harm that makes it unavoidable, and the only choice a profession really has is whether to build it a little ahead of that moment or a long way behind it. We are, for now, still ahead of it. That will not last. The question each organisation should be asking, in the quiet before the framework is written for it, is a plain one: when the reckoning comes — and on the evidence of every prior technology, it will — do we want to have been the ones who saw it coming, or the ones explaining, afterwards, why we did not?