

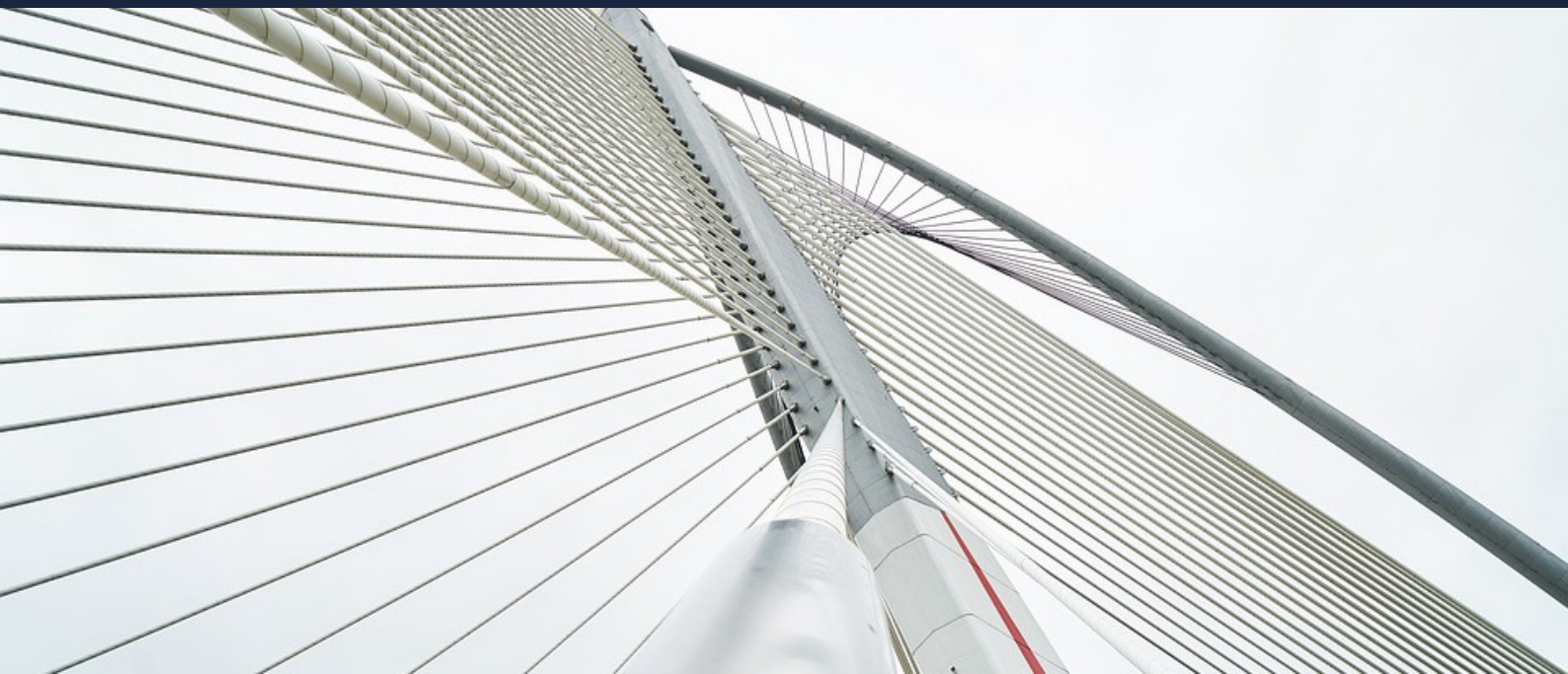
BEYOND AUTOMATION ANXIETY

The Verification Economy — When Producing Became Cheap and Checking Became the Job

On Apprenticeship Economics, Machine Production, and the Quiet Depletion of Organisational Judgement

Giovanni Leonardi

September 2025



Contents

1. Executive Summary	3
2. The Review That Revealed the Gap	3
3. The Hidden Curriculum of Production	4
4. When the Training Ground Disappears	5
5. The Demography of Judgement	5
6. The Counterargument and Its Limits	6
7. What Rebuilding Might Look Like	7
8. The Responsibility We Have Not Yet Named	8

The Verification Economy — When Producing Became Cheap and Checking Became the Job

Leadership · No. 107
September 2025

"They were reviewing work they had never learned to do."

1. Executive Summary

The industrialisation of content production through generative AI has triggered a less visible but more consequential shift: the work that matters most in organisations is moving from creation to verification, from drafting to judging. This essay examines the structural consequence that few organisations have yet confronted — that the apprenticeship pipeline which once turned producers into evaluators depended on production itself as the training ground, and that ground is disappearing. As machine output floods knowledge-work workflows, organisations are drawing down their reserves of human judgement without replenishing them. The seniors who can still distinguish good from adequate were formed by decades of doing the work that juniors will never do. What follows is not an argument against automation, but an exploration of the skill-formation economics that automation has disrupted, and the deliberately unglamorous capability programmes that might begin to restore them.

2. The Review That Revealed the Gap

A quality assurance lead in a professional services firm described a moment that had been troubling her for months. Her team of six analysts had been reviewing AI-generated client deliverables — reports, models, summaries — for the better part of a year. They were fast, diligent, and had mastered the review rubrics. But when she began tracking the errors they caught against the errors that actually reached clients, a pattern emerged: they were excellent at catching formatting inconsistencies, structural gaps, and factual errors verifiable against source data. They were poor — conspicuously, systematically poor — at catching errors of judgement. Misframed problems. Technically correct analyses that answered the wrong question. Recommendations that made sense in isolation but ignored the political dynamics of the client organisation.

These were precisely the errors that the firm's partners, with twenty or thirty years of production behind them, caught instinctively. The analysts, most of whom had joined after the firm adopted generative tools, had never produced a client deliverable from a

blank page. They had never sat with an ambiguous brief, made the hundred small judgement calls that shape a report's argument, or experienced the slow education of having a partner dismantle their draft and explain why it missed the point.

They were reviewing work they had never learned to do.

*"We have trained a generation of quality controllers
who do not understand quality."*

The quality assurance lead's conclusion was blunt, and it carries further than she perhaps intended. It is not a statement about training budgets or onboarding programmes. It is a statement about the economics of professional formation — and those economics have shifted beneath every knowledge profession simultaneously.

3. The Hidden Curriculum of Production

The insight buried in this observation is not really about artificial intelligence. It is about apprenticeship economics — the centuries-old model through which professions reproduce their judgement.

Every knowledge profession has operated on the same implicit bargain. Juniors enter and produce: they draft contracts, write code, prepare analyses, compose reports. The production is the point, but not for the reason organisations usually articulate. The ostensible purpose is output — the firm needs the work done. The actual purpose, visible only in retrospect, is formation. Through producing, juniors absorb the tacit knowledge that separates competent practitioners from excellent ones: the sense of what a good argument feels like before you can articulate why; the instinct for which simplification is acceptable and which is dangerous; the ear for a sentence that sounds authoritative but says nothing.

This tacit dimension of professional skill has always resisted codification. We have called it "experience," "instinct," "good judgement" — labels that gesture toward something real but resist the kind of specification that would make it teachable in a classroom. The apprenticeship model solved this problem not by codifying the tacit but by ensuring enough repetition, enough exposure, enough corrected failure that the tacit was absorbed through practice. A junior solicitor who has drafted three hundred contracts has internalised constraints and patterns that no training course could convey. A software engineer who has debugged a thousand production incidents has developed a diagnostic intuition that no certification can substitute.

The critical feature of this model — the feature now under threat — is that production was the mechanism of formation, not merely its context. Juniors did not learn judgement and then produce. They produced, and judgement emerged as a byproduct of the producing.

4. When the Training Ground Disappears

Generative AI has not eliminated production. It has made it cheap enough that the economics of who does it have shifted. When drafting a report takes a machine forty seconds rather than a junior analyst three days, the rational organisational response is to let the machine draft and redeploy the analyst. The redeployment, almost universally, is to review, verification, and quality assurance.

The logic is impeccable in the short term. Output increases. Seniors are freed from routine production oversight. Juniors, now positioned as reviewers of machine output rather than producers of human output, can process a higher volume of work. Every efficiency metric improves.

What does not improve — what in fact begins to deteriorate in ways that are difficult to measure and easy to ignore — is the formation pipeline. The juniors who are reviewing AI-generated work are not undergoing the apprenticeship that would eventually make them capable of the judgement their review role demands. They are performing a function — checking, verifying, approving — without having traversed the developmental path that makes the function meaningful.

In software engineering, the pattern is already visible. Teams that adopted AI coding assistants early report a specific and recurring problem: junior developers who can review and approve machine-generated code with reasonable competence when the code follows familiar patterns, but who lack the architectural intuition to recognise when the pattern itself is wrong. They can verify that the code works. They cannot judge whether it should have been written differently. The distinction between verification and judgement — between confirming that something is correct and assessing whether it is good — is precisely the gap that production experience once closed.

In legal practice, a parallel dynamic is emerging. Junior lawyers reviewing AI-drafted contracts can identify missing clauses and formatting errors efficiently. What they struggle to assess is whether the contract's overall structure of risk allocation is appropriate for the commercial relationship it governs — a judgement that senior practitioners make by drawing on hundreds of negotiations they personally conducted. The juniors have access to the same precedent databases, the same checklists, the same style guides. What they lack is the residue of having sat across a table and felt the negotiation shift.

5. The Demography of Judgement

Organisations are not confronting this as a systemic problem because the consequences are masked by a demographic buffer. The seniors — partners, principals, lead engineers, editorial directors — who built their judgement through decades of production are still in place. They still catch the errors that junior reviewers miss. The firm still functions.

But this buffer is finite and non-renewable on current terms. The practitioners who entered knowledge professions in the 1990s and early 2000s, who served full apprenticeships in the old production-heavy model, are in the second half of their careers. They will retire, or leave, or simply burn out from carrying an increasing share of

the judgement load as the ratio of reviewers to evaluatively competent reviewers deteriorates. When they do, the generation behind them — formed in a production-light, review-heavy environment — will inherit the evaluative responsibilities without the evaluative capacity.

The demographic mathematics are stark when stated plainly. In a large consultancy, the partners who joined before widespread AI adoption might have fifteen to twenty years of production-formed judgement behind them. The managers who joined afterward have three or four years of review experience and access to better tools. The partners carry the accumulated judgement of hundreds of engagements in which they personally wrestled with ambiguity, made consequential choices, and learned from the results. The managers carry fluency with AI tools, review protocols, and efficiency frameworks — genuine and valuable skills, but categorically different from the evaluative depth that their future roles will demand.

We are consuming our judgement reserves without replenishing them. The depletion is invisible in quarterly reporting. It will be starkly visible in the quality of decisions a decade from now.

6. The Counterargument and Its Limits

The obvious objection is that judgement can be taught directly — that we need not rely on the indirect, inefficient, and frankly wasteful mechanism of having juniors produce work that a machine could produce faster. Why not design training programmes that develop evaluative skill deliberately? Why not create simulations, case studies, structured review exercises that build judgement without requiring the slow accumulation of production experience?

This objection deserves serious engagement, because it is partly right. The old apprenticeship model was never designed to develop judgement. It did so as a byproduct, which means it did so unevenly, unsystematically, and with enormous waste. Many practitioners served long apprenticeships and never developed particularly good judgement. The model worked at the population level — enough people, given enough production experience, developed enough evaluative skill to staff the senior ranks — but it was brutal in its inefficiency and opaque in its mechanisms.

A deliberate programme of judgement formation could, in principle, be more efficient and more equitable. It could make the tacit explicit, accelerate the development of diagnostic intuition, and open evaluative roles to people who were excluded from the old production-heavy pipeline.

But two difficulties qualify this optimism considerably.

The first is that we do not yet know how to do it well. The tacit knowledge embedded in professional judgement has resisted codification not because nobody has tried, but because the knowledge itself resists the kind of articulation that teaching requires. Knowing that a legal argument "doesn't feel right" before you can say why is a real cognitive capacity, and we do not have reliable methods for transmitting it outside the context of practice. The research literature on expertise development — from the Dreyfus model of skill acquisition through Ericsson's work on deliberate practice — offers useful

frameworks but nothing approaching a proven curriculum for evaluative judgement in professional contexts. The gap between "we understand that expertise develops through staged experience" and "we can design a programme that reliably produces it" remains wide.

The second difficulty is organisational rather than pedagogical. Even where elements of deliberate judgement training are feasible, they require investment in something with no immediate return. The apprenticeship model generated formation as a byproduct of billable production — it cost the organisation nothing explicit. A deliberate programme costs time, senior attention, and tolerance for the short-term inefficiency of having people practise judgement on work that could be processed faster without them. In an environment where every efficiency gain from AI adoption is celebrated and measured, the case for deliberately slowing down to rebuild a capability that has not yet visibly failed is an extraordinarily difficult case to make to a management board.

7. What Rebuilding Might Look Like

Yet the difficulty of the case does not diminish its urgency. Organisations that recognise the problem — and a growing number do, even if they lack a shared vocabulary for it — are beginning to experiment with approaches that attempt to rebuild what the apprenticeship pipeline once provided incidentally.

The most promising share certain characteristics. They involve exposure to real ambiguity rather than sanitised case studies. They require the exercise of judgement under conditions where the answer is genuinely uncertain and the consequences of error are felt, not hypothetical. And they make the evaluative process visible and accountable in ways that the old apprenticeship, with its informal and often arbitrary feedback, never managed.

Three patterns recur across sectors:

- The error museum. A curated, anonymised collection of consequential mistakes, maintained not as a compliance exercise but as a pedagogical resource. The errors are selected not for severity but for instructiveness: cases where technically correct work led to poor outcomes because the framing was wrong, the assumptions went unchallenged, or the context was misread. Juniors study these not to learn rules but to develop the pattern recognition that rules cannot capture. The error museum works because it provides, in compressed form, something close to the slow accumulation of "cases seen" that production experience once supplied.
- Accountable sign-off with teeth. The deliberate restructuring of approval processes so that reviewers bear genuine professional responsibility for the work they approve, not merely procedural liability. When a junior reviewer's name on a deliverable carries real consequences — when they will be asked, in the post-engagement review, to explain not just what they checked but what judgement calls they made and why — the review ceases to be a quality-control function and becomes a judgement-development exercise. The stakes transform the cognitive activity from pattern-matching to genuine evaluation.

- Structured production intervals. Deliberate periods in which juniors produce work from scratch, without AI assistance, under close senior mentorship. These intervals are explicitly unproductive by efficiency metrics. Their purpose is formative: to force the engagement with ambiguity, the sequence of consequential choices, the experience of having one's thinking dismantled and rebuilt, that the old apprenticeship provided as a matter of course. Some firms have begun embedding these as a formal career-stage requirement, treating them as the professional equivalent of a medical residency — costly in the short term and non-negotiable for long-term competence.

These approaches share a common feature that distinguishes them from conventional training: they do not attempt to teach judgement as a body of knowledge. They attempt to create the conditions under which judgement develops — conditions that the old production-heavy model created accidentally and that the new production-light model has inadvertently removed.

8. The Responsibility We Have Not Yet Named

The deeper argument of this essay is not that organisations should resist automation or slow their adoption of generative tools. The efficiency gains are real and, in most cases, genuinely beneficial. The argument is that the adoption has carried a cost — to the formation of professional judgement — that was neither anticipated nor accounted for, and that the cost compounds silently.

This is, at its core, an organisational responsibility. Individual practitioners cannot solve it by choosing to produce more slowly. Markets will not solve it, because the degradation of judgement quality manifests too gradually and too diffusely to generate a corrective price signal. Regulators are unlikely to solve it, because the problem is not one of compliance but of capability. It falls to the organisations themselves — the firms, the departments, the professional bodies — to recognise that the pipeline which once formed their evaluative capacity has been disrupted, and to invest deliberately in something to replace it.

The verification economy is not a temporary transition. The proportion of knowledge work that consists of reviewing, evaluating, and taking responsibility for machine-generated output will only increase. The question is not whether organisations will need people who can judge — it is whether they will still have them.

That investment will be unglamorous. It will look like wasted time to anyone measuring productivity by volume of output processed. It will require senior practitioners — the very people whose time is most expensive and most in demand — to spend hours mentoring juniors through work that a machine could complete in minutes. It will require management boards to think in timeframes that quarterly reporting actively discourages.

But the alternative is to continue drawing down the reserves — to keep consuming the judgement that a previous generation's apprenticeships laid down, without replenishing it. And the nature of reserves is that they feel inexhaustible right up until the moment they are gone.