The Scaffolding of Intellect

Method Over Magic: How Organisations Tame a Non-Robust Technology

Conor Flynn | Trinzo | July 2026

Executive Summary

The two papers before this one were diagnoses: the first argued that AI amplifies what an organisation already has rather than manufacturing what it lacks; the second, that the most common human-in-the-loop workflow corrodes the very capability that determines whether any of this works at all.

This paper is the answer to both, and it is deliberately the most assertive of the three, because a series that spends two papers describing a disease owes its reader a treatment. The argument is that AI's value is a function of methodological sequencing: the organisations that win will be the ones that erect scaffolding around a probabilistic engine rather than waiting for a magic button. I mean the word scaffolding precisely, and I intend to earn it by the end: structure built around the work rather than instead of it, cheap relative to what it enables, and designed to let ordinary people build safely at heights that would otherwise be lethal.

Four arguments carry the paper, and I want to be clear at the outset about which are mine and which are not, because a paper about verification should survive its own standard.

The first is commercial. The models are going to commoditise. When every competitor holds the same engine, method becomes the last available differentiator. This is an application of Perez's framework, not a discovery.

The second is that the failure most organisations are experiencing is a mismatch rather than a malfunction. They are running a probabilistic engine inside a deterministic expectation, and the property that causes the trouble is narrow: the output does not signal its own error. Fluency has been decoupled from accuracy, and most of what follows organisationally, the rubber-stamp reviews, the workslop, the metrics that look healthy while the substance rots, follows from that single fact.

The third is that verification, not generation, is now the binding constraint on AI-assisted work, and that for fluent-and-wrong output the ordinary economics of checking invert. I arrived at this through practice; a converging literature has arrived at the same place, and I present it as synthesis. What I add is operational: a placement rule for gates, a test for when human oversight is real, and an account of which professions are most exposed.

The fourth is that account of exposure, and it is the claim I have not found elsewhere. The disciplines with explicit verification traditions, engineering, medicine, audit and law, built them because their failures were expensive and visible. The disciplines that run on plausibility never had to build that immune system, and the deficiency is now being tested for the first time.

Where I draw on my own practice, I will say so. Where the evidence cuts against me, that will be visible too.

1. Too Much Airplane for One Man to Fly

On the thirtieth of October 1935, at Wright Field in Ohio, Boeing's Model 299, the aircraft a journalist had already named the Flying Fortress, stalled at three hundred feet and burned, killing two of the five men aboard, among them one of the most experienced airmen in the service. The investigation found no fault with the design. The gust locks had been left engaged. It was the failure of a capable person to hold every necessary step in his head while operating something more complex than anything he had operated before, and one newspaper produced the verdict that outlived the aircraft: it was too much airplane for one man to fly.

The Army did not order a simpler aeroplane, and it did not wait for better pilots. A group of test pilots produced a checklist short enough to fit on an index card, and with that structure in place the unflyable aircraft became one of the most consequential machines of the twentieth century (Gawande, 2009).

The story is well worn. Gawande built a book on it, and half the management literature has borrowed it since. I use it anyway, because it remains the cleanest instance on record of a capability crisis answered by method rather than by a better machine or a better operator, and that is precisely the shape of the present moment.

Notice what the index card was, because the whole of this paper is in it. It was not a limitation on the aircraft, and it was not a training course for the pilot. It was scaffolding: structure external to both the machine and the operator, almost embarrassingly cheap relative to what it enabled, and load-bearing from the day it went up. The aircraft supplied the capability. The card supplied the conditions under which the capability could be used without killing anyone. Most organisations today have been handed too much airplane. Most are still trying to fly it on instinct.

Diagnosis is the easier half of this subject and it has been overserved. There is a substantial and largely correct literature explaining why enterprise AI is underperforming, and a much thinner one explaining what to do instead that does not collapse into vendor selection or platitude. This paper is my attempt at the second thing.

2. Why This Matters Now: The Commodity Argument

Before the method, the commercial case, because a chief executive is entitled to ask why workflow design should be a board-level concern rather than something delegated three layers down.

Perez's account of technological revolutions describes a consistent migration: during the installation period, value concentrates on those building the infrastructure; during the deployment period that follows, it migrates to those who use it well (Perez, 2002). The first paper used that framework to read the financial cycle. Applied here, it produces a straightforward prediction: the models will become commodities. The signs are already visible in falling cost per token, in open-weight models closing the gap on an increasing

range of tasks, in multiple providers at rough parity at the frontier, and in switching costs that remain remarkably low for anything this strategically significant.

Commodity does not mean unprofitable. Electricity generation is a commodity and utilities are durable businesses. What commoditisation destroys is pricing power and excess return. The likely destination is that frontier labs earn utility-like returns on extraordinary capital: materially worse than what is currently priced in, but not zero, and I would not want to be read as predicting collapse.

The strategic consequence is what matters. If the model is the differentiator, the correct response is to buy a better model, and methodology is a nice-to-have. If the model is a commodity, method is the only thing left to compete on, because your competitor is holding the identical engine.

There is a serious objection: proprietary data and distribution are the obvious rival candidates for durable advantage. My answer, offered as reasoning rather than evidence, is that data and distribution are positional. They can be bought, replicated, regulated away or disintermediated. Method is reproductive. It compounds through people, improves with use, and is difficult to acquire by purchase, because what would be acquired is a set of habits rather than an asset. I hold this with reasonable but not complete confidence.

The electrification history the first paper told at length makes the same point in one sentence here: the durable returns went not to those who generated the power but to the manufacturers who redesigned the factory floor around it (David, 1990). The advantage lived in the structure, not the engine, and I see no reason to expect this cycle to distribute its rewards differently.

3. The Failure Mode: Fluency Decoupled from Accuracy

The dominant expectation inside organisations is what I call the Magic Button: one prompt in, a finished corporate asset out. The model is treated as a deterministic oracle, budgets and timelines are built on that assumption, and the disappointment when it fails to materialise is read as evidence that the technology does not work.

The technology works. What does not work is running a probabilistic engine inside a deterministic expectation. The same prompt, run twice, can produce materially different output; the system has no stable notion of the truth of what it produces, only of what is likely to come next. This is not a defect awaiting a patch. It is what the thing is, and the mismatch accounts for a great deal of the gap the MIT NANDA study measured when it reported 95% of generative AI pilots delivering no measurable financial impact (Challapally et al., 2025), a figure I use here, discounted for the reasons the first paper set out, as one signal among several.

It is worth being honest about where the Magic Button expectation came from: it was sold. The most confident claims about the imminent end of knowledge work have been made by people holding the largest positions in the outcome, into a capital-raising environment, and should be read accordingly.

What the Magic Button actually produces, at scale, is slop, and the evidence has moved from anecdote to measurement. Graphite's analysis found the share of newly published web articles primarily generated by AI crossed fifty percent in early 2025 and has plateaued there (Graphite, 2026). Note that there are known issues with the methodology behind this analysis. Inside firms, research from BetterUp Labs and Stanford's Social Media Lab gave the phenomenon a name, workslop, and put numbers on it: forty percent of employees reported receiving low-substance AI-generated work product in the previous month, each incident consuming close to two hours to untangle (Niederhoffer et al., 2025). In software, vibe-coding is the same phenomenon in a domain where consequences are legible.

Notice what these findings share. In every case the failure is not that the machine produced nothing. It is that the machine produced something that looked right.

In almost every other tool an organisation uses, failure announces itself. The build fails; the machine stops; the number is visibly wrong. Confidence and correctness travel together closely enough that a competent person can triage on surface signal, which is why organisations have run on review-and-approve workflows for a century. The first paper observed that generative AI has made plausibility free, and that plausibility was the signal we relied on to detect competence. This paper needs the precise version of that observation, because everything that follows is built on it. Probabilistic behaviour is not itself the problem; systems can be probabilistic and reliable. The problem is narrower and worse: the output does not signal its own error. The failure mode is not an error message but a well-formed paragraph that happens to be wrong. Workslop does not look like slop. That is why conscientious people wave it through, and why asking people to review more carefully has no purchase whatsoever.

The second paper traced what Jevons's paradox does to cognitive work, and the consequence lands here: as the marginal cost of generating a document approaches zero, volume expands, and because the outputs do not announce their errors, the burden of sorting the good from the merely plausible falls entirely on human attention, which has not become cheaper at all.

The bottleneck has moved. It is no longer whether we can produce it. It is who can tell whether it is any good.

4. What the Engine Actually Does

If the expectation is wrong, the useful question is what the engine is for. In my own practice the value divides along one axis, and most failed deployments are explained by pointing the tool at the wrong side of it.

I begin with a concession stronger than the one usually made in papers of this kind. On clear, bounded tasks with relatively provable answers, machines can produce better and more consistent output than humans. Not merely faster or cheaper. Better. Arguments about AI that begin by denying this are not worth reading. The refinement that matters is that a task is rarely born clear and bounded. It becomes so because somebody did the contextual work of specifying it, defining what counts as a correct answer, and deciding

where the boundary belongs. That specification work is judgment of exactly the kind machines are poor at. The human does not disappear from the bounded-task category; the human moves upstream into the bounding, which is precisely where this paper is going to put the gates.

The engine is very good at macro-aggregation: processing volumes no person could hold, producing the first structured view of a large and messy territory. It is good at personalisation at scale, extending the reach of work that is already good. And it is very good at defensive automation, the most underrated category: eliminating work that should never have needed doing. The spam filter is among the most successful applications of machine learning in history precisely because nobody framed it as replacing a person, and in an environment where synthetic volume expands faster than human attention, the capacity to filter intelligently will matter more than the capacity to generate.

Core execution is where the engine breaks: work that is high-stakes, hyper-contextual and intolerant of error. The standard explanation, that models lack context, is half right in a way that has become unhelpful, because it invites the response that context windows are getting longer. The more precise failure is about weighting rather than volume. Early in a piece of work one set of facts is load-bearing; several stages later a different set is, and the first has become noise. Humans re-weight continuously and mostly without noticing, because they track what the work is for. Models treat context far more flatly, and anyone who has run a long multi-stage piece of work through one has watched it solve the problem it was given four steps ago rather than the one in front of it.

The durable version of this claim, the version that survives longer context windows, retrieval and memory architectures, is structural. A model can only consume what has been written down. The weighting of what matters in a real situation is a function of stakes, relationships, institutional history and consequence, most of which was never recorded anywhere, and much of which could not have been. Polanyi's tacit dimension has quietly underwritten all three of these papers (Polanyi, 1966), and it does not depreciate with the next model release.

Bletchley Park is the standard historical reference on this point, and it is usually deployed badly, as a story about a machine that broke a code. The actual operation was an industrial-scale human process with a machine embedded in one stage. The bombe did not read German and did not produce intelligence; it narrowed an impossible search space to a set of candidates humans could test. Everything that made the output valuable, the cribs, the traffic analysis, the translation, the assessment of what a decrypt meant and what could safely be done with it, happened around the machine, inside a strict sequential structure with human judgment placed where judgment was required. Nobody expected the machine to hand them a finished product, and the whole apparatus was built on the assumption that it would not.

5. Humans Are Load-Bearing

The prediction implied by the Magic Button is that as capability rises, headcount falls. The early data does not support it.

The most useful evidence so far comes from Ramp's Economics Lab with Revelio Labs, analysing spend and workforce data across more than twenty-one thousand US companies: firms in the highest tier of AI adoption maintained employment roughly 10.2% higher than non-adopters over the following twenty-four months, and their entry-level workforce share rose (Kharazian, Simon & Stevens, 2026). PwC's 2026 Global AI Jobs Barometer points the same way (PwC, 2026). The caveat matters more than the headline: heavy adopters are disproportionately fast-growing, well-capitalised firms that were hiring anyway, and twenty-four months is not long enough for industry-wide conclusions. What the data supports is the narrower claim I need: there is no visible evidence yet of the workforce contraction that was widely predicted, and some evidence of the opposite, which is worth knowing precisely because so much strategy has been built on the assumption of contraction.

The second pattern is the hire-back, with Klarna the canonical case, examined at length in the first paper and needing only its lesson here: the deployment optimised for cost when the firm's actual advantage was frictionless experience, the metrics looked excellent because they measured the wrong thing, and by mid-2025 the company was rehiring the capability it had celebrated replacing.

The third pattern is the cleanup economy, the most telling because it is a market forming in real time around a gap. Marketplaces now exist specifically to repair vibe-coded software; agencies have emerged whose business is rewriting AI-generated copy and untangling broken AI-assisted builds. Firms are paying, after the fact and at premium rates, for exactly the human structuring work they believed they had automated away.

Read together, these patterns say something specific. Value is not migrating away from humans. It is migrating toward the humans who impose structure. And if the scarce and appreciating asset is structured human judgment, then unreflective automation is removing precisely the capability that is becoming most valuable, which converts the second paper's cultural argument into a strategic one.

Which raises the question this paper exists to answer. If humans are load-bearing, what exactly are they supposed to do?

6. Verification Is the Binding Constraint

Mechanism first, because "break the work into stages and put a human at each boundary" sounds like process consulting and deserves to be dismissed as such unless there is something underneath it. There is.

The arithmetic of compounding error. Take any multi-step piece of knowledge work and treat each step as having some probability of being carried out correctly. Run unchecked, one step feeding the next, the reliability of the whole is roughly the product of the reliability of the parts: a chain of ten steps at ninety percent delivers something in the region of thirty-five percent, and because the output does not announce its own error, the thirty-five percent arrives with the same fluent confidence as the ninety. Place a checkpoint between steps, and an error caught at the boundary does not propagate into everything built after it.

Two honest caveats, because the clean version of this arithmetic proves too much. It assumes the steps fail independently, which they do not; errors cluster around a bad framing the way cracks cluster around a flaw. And it assumes the gates catch what passes through them, which they will not; every checkpoint has its own miss rate. The claim that survives both caveats is the one I need: gating changes the decay from multiplicative toward additive. It does not make it zero, which is why the engineering tradition this borrows from never relies on a single barrier. Von Neumann posed the general problem in the early 1950s under a title almost uncomfortably apt for our situation: Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components (von Neumann, 1956). James Reason's defence-in-depth model, in which no single control is expected to be sufficient and reliability comes from the arrangement of imperfect ones, is the same idea worked out for organisations (Reason, 1997). The checklist is not a productivity habit. It is an error-containment structure, imperfect by design and reliable in aggregate.

There is a working posture underneath all of this that deserves to be stated as a principle. Design every AI-assisted workflow on the assumption that the machine has already failed somewhere inside it, and treat the only open question as where the failure gets found. This is not pessimism about the technology; it is the posture aviation takes toward engines and medicine takes toward diagnoses. Its practical force is that it reverses the default: an error is presumed to exist until a gate demonstrates otherwise. Expect the machine to fail, and the sequence stops being overhead and becomes the search.

What is actually capped. I have claimed in various forms for a couple of years that the quality of AI output is capped by the critical-thinking capacity of the operator. That is imprecise, and I want to correct it publicly. AI can plainly produce work exceeding what a given operator could produce unaided; if quality were capped by the operator, that could not happen. What is capped is the operator's ability to recognise quality. You cannot accept, reject, improve or safely deploy what you cannot evaluate. The ceiling is not on generation. It is on verification.

The inversion. In most domains, verifying is cheaper than producing: easier to check a proof than find one, easier to taste the soup than cook it. That asymmetry is the unexamined assumption under every human-in-the-loop diagram I have been shown, and it is why review-and-approve has been a sensible division of labour for as long as organisations have existed. For fluent-and-wrong output, the asymmetry inverts. When plausibility is high and correctness is unknown, genuine verification means reconstructing enough of the underlying reasoning to establish whether the answer is right, which is most of the work of producing it, plus the burden of resisting a text that is actively signalling its own competence. Checking becomes more expensive than doing.

I should be accurate about the provenance of this claim, because when I first formulated it from practice I believed it was mine, and the literature says otherwise. Over the past eighteen months a converging body of work has arrived at the same inversion from several directions: from software engineering, where review effort now measurably exceeds production effort (Garousi, 2026); from scientific workflows, where generation costs have collapsed while verification costs have not (Ma, 2026); from the economics of AI, in

Acemoglu's distinction between checkable outcomes and context-sensitive judgment (Acemoglu, 2025); and from MIT Sloan's work tying realised AI value to output verifiability. The inversion itself should now be regarded as an emerging consensus. The contribution of this paper is what follows operationally from taking it seriously: where the gates go, what makes one real, and who is most exposed.

The inversion also completes the second paper's account of the reviewer trap in a way I could not manage at the time. Bainbridge showed that monitoring roles erode the skills they depend on and that vigilance degrades on reliable systems (Bainbridge, 1983); Kahneman explains why fluent text recruits fast, automatic cognition when the task requires slow, deliberate effort (Kahneman, 2011). Both are true. But the deeper reason the arrangement fails is economic. You have handed the human the more expensive job, in less time, with worse information, and then measured them on throughput. Of course it collapses into approval. It was never affordable as designed.

The evidence that cuts the other way. If output is bounded by the operator's judgment, you would expect AI to widen the gap between strong and weak performers. Several of the most-cited studies find the opposite. The Harvard Business School and BCG field experiment on consultants found the largest gains accruing to below-average performers, compressing the distribution (Dell'Acqua et al., 2023); Brynjolfsson, Li and Raymond found the same shape among customer support agents (Brynjolfsson, Li & Raymond, 2023). On its face this challenges everything above.

The reconciliation runs along the axis this section has been building. Both studies examine work where the quality of an answer is easy to establish: a support issue is resolved or it is not; a consulting task in a controlled experiment has a gradeable output. Where output is readily verifiable, AI compresses performance, because verification is cheap for everybody and the tool supplies what the weaker operator was missing. Where output is hard to verify, in the high-context, tacit core execution work, the operator's judgment is the only available check, and I expect the distribution to widen. Notably, in the BCG study the tasks designed to sit outside the model's frontier saw assisted consultants perform worse than unassisted ones.

The defensible version of the claim is therefore: AI compresses performance on verifiable work, and I predict it widens performance on unverifiable work. The asymmetry in that sentence should be flagged. The compression half is well evidenced. The widening half is, so far, inference: the published evidence shows gains becoming variable, vanishing, or reversing on ambiguous and judgment-heavy tasks, which is divergence in the predicted direction rather than demonstrated widening, and recent modelling work supports the mechanism without settling it (Huang, Xiao & Vishnoi, 2026; An, 2025). It is a testable claim, and I would rather it be tested than believed.

It is also immediately useful, because it tells an organisation where to concentrate attention. The tasks where AI most flatters your weakest performers are the ones to worry about least. The trouble concentrates precisely where you cannot easily tell how you are doing, which is, inconveniently, where most of the value in a professional firm sits.

7. Two Paths, Two Structures

Most of the conversation about getting value from AI collapses two different problems into one. There are at least two paths to value, they behave differently, and they require different scaffolding. Confusing them is among the most expensive mistakes an organisation can make in this territory, so this section is prescriptive by intent.

The first path is scaling the technology: the organisational problem of building reliable systems from an unreliable component, through gated sequence, error containment and defined handovers. It is standardised by nature; its structures must be repeatable, auditable and transferable, because the point is that the output survives contact with an organisation rather than depending on a single practitioner. It is an engineering problem, and the operating procedure described in the next section is an artefact of it: it exists so that a regulated client's output does not depend on which practitioner happened to be at the desk that week.

The second path is amplifying the individual, which is what I see working most consistently in my own practice. It is personalised, agile and specific: a particular person applying particular judgment to a particular situation. The structure that makes it work is not a process document but something closer to a craft discipline. My own working method is an artefact of this path: the sequence I run analytical work through, frame before generating, generate against the frame, attack the output, reconcile, was tuned to my strengths and failure modes over three years, and it transfers badly. When I have tried to hand it to colleagues as a procedure, what transfers is the steps; what does not transfer is the judgment about when to abandon them. That is not a failure of documentation. It is the nature of craft.

In both cases, imposed structure is what brings the value. But if structure explains value everywhere it explains nothing, so the claim needs its specification: the two paths require different structures, and each fails when built with the other's materials. Organisations routinely try to solve the second path with the instruments of the first, issuing prompt libraries and mandated workflows to individuals whose value came from personalised judgment, and wonder why adoption is sullen. The converse error is rarer but more expensive: assuming that because a few individuals are producing extraordinary work, the organisation has a capability, when what it has is several talented people and no transferable method. The test is worth applying in its bluntest form: if your best AI-assisted performer resigned tomorrow, what would remain? If the answer is a licence and a folder of prompts, you are on the second path while reporting to the board that you are on the first.

Where does each structure live? At the level of the organisation, structure is standardised: policy, governance and shared gates. At the level of function and role, the two paths meet: a role has a characteristic distribution of task verifiability, and this is where the decision gets made about which work is bounded enough for the machine to lead. Most of the practical design work happens here, and it is the level most often skipped, because it is nobody's obvious job. Assign it to somebody. At the level of the individual, structure is personalised

and has to be; this is where the second path operates, and where imposed uniformity does the most damage.

This distinction also supplies part of my answer to the strongest objection to this whole paper, which I take up in Section 10: the bitter lesson says nothing about the return on personalised human method, because that is not something a general method can eat.

8. Designing the Gates

If the mechanism is error containment and the scarce resource is verification, the design of work follows from two questions: where do the gates go, and what has to happen at one for it to be real? Both are answered from the posture stated in Section 6. The machine is assumed to have failed somewhere; gates are where you go looking.

Where gates belong. Placement stops being a matter of taste. A gate belongs wherever the cost of a propagated error first exceeds the cost of checking: at the boundaries where a decision becomes load-bearing for everything after it. The context-weighting problem from Section 4 gives a second placement rule: if models are poor at recognising when the salience of context shifts, gates belong precisely where it shifts. A gate is where re-weighting happens, where a person establishes what now matters, what has stopped mattering, and what has entered the picture that was not there before. This is among the most valuable things a human does in an AI-assisted workflow and it is almost never designed for explicitly.

Those two rules generate the stages rather than the other way round. In analytical work the expensive-to-reverse commitments are the framing of the question and the selection of evidence, so the sequence gates early and often; in execution work they sit in the decomposition of the problem and the assumptions embedded in scheduling and risk, so gates fall there. A different domain would generate different boundaries from the same principles, and what I would defend is not any list but the rules that produce it.

A related principle matters as much as placement: iterate within a stage rather than across the whole. The instinct with a capable model is to one-shot the entire task and fix what comes back, which is the most expensive possible way to work: it maximises the length of the unchecked chain and defers all verification to the point where verification is hardest. Looping inside a bounded stage keeps every act of checking small enough to actually perform.

What makes a gate real. The obvious objection comes from my own previous paper. A phase gate staffed by somebody who glances at output and waves it through is the reviewer trap with better paperwork. Install six of those and you have built six opportunities to rubber-stamp, plus an audit trail implying a diligence that did not occur. That is worse than nothing, because the organisation now feels safe.

The governance literature has described this failure extensively; what it has mostly not supplied is an operational test, so here is the one I use. A gate is real when passing it requires the human to produce something the machine did not. A decision, a rejection, a reframing, a constraint, a piece of context that exists only inside the firm. If the human's

contribution can be discharged by approval, it is theatre, and it will decay into theatre within weeks regardless of what the process document says. If passing requires generation, if the work genuinely cannot move forward until a person has put something of their own into it, then the human is in the work, in the sense the second paper demanded, rather than merely near it.

This criterion is not hypothetical; it is running in my own delivery work. In a training content programme for a regulated life-sciences client, the operating procedure runs the work through five gated phases, from brief and research through to final assembly. Drafts are generated with one model and then attacked by a second running structured adversarial personas, an auditor, a technical specialist, a novice end-user, before any human reviewer sees them. Every piece of machine feedback is then classified by a person as valid, already addressed, or wrong, because AI reviewers flag issues that are already correct and miss ones that are not. And before anything goes to subject-matter experts, the practitioner must personally verify every high-risk regulatory claim against the primary source, the actual regulation rather than a summary of it, and log each check in a verification record that travels with the draft. One AI model confirming another AI model's claim is pattern matching, not verification. The record is the gate artefact: something the machine did not and could not produce, and the reason the gate cannot decay into approval.

This is also a compliance specification. The EU AI Act requires meaningful human oversight for AI systems in high-risk contexts (European Parliament and Council, 2024). A name in an approval workflow will not satisfy that standard; a documented, substantive human contribution at defined points will, and the gate criterion happens to be the cheapest way to produce one. Method and compliance pointing in the same direction does not happen often enough to waste.

Gates are handovers. The first paper argued that the central unsolved problem in applying AI at scale is trust in handover: if I do not know what you asked, what you verified and what the machine did, I cannot build on your work without rebuilding it. I did not see, when I wrote that, that it is the same problem described here, with the same solution. A gate is a handover point. The artefact a real gate produces, the decision taken, the assumption made explicit, the thing rejected and why, is precisely the context a colleague needs to extend the work rather than reconstruct it. The scaffolding around one person's work turns out to be the bridge to the next person's.

This explains something I had observed for years without being able to account for: individuals and small teams generating remarkable value with these tools, and that value very rarely scaling. Small teams work because verification stays cheap: shared context, visible work, everyone able to check everyone. Scale breaks it because verification cost grows faster than headcount; every handover adds a point where somebody must establish quality without having been present when the work was made. The electrification lag of Section 2 is usually read as a counsel of patience, and I think that misreads it. The thirty years were spent discovering that the factory floor had to be redesigned. The redesign is the work, and whoever does it first takes the gain.

The gates will be gamed. The second paper described what happens when a measure becomes a target, and it would be careless not to apply that to my own prescription. Install six phase gates, measure the organisation on gates passed or cycle time, and gate artefacts will become a genre of writing rather than a record of thought; the dashboard will look immaculate while the substance rots, the pattern I have spent two papers warning about, reproduced inside the fix.

The protection is structural and follows from the gate criterion. When the work cannot proceed without something only a person can supply, the incentive to game largely collapses, because there is nothing to game. You cannot fake a decision the next stage depends on. What you can fake is an approval, which is why approval-shaped gates are the ones that rot. Measure the organisation on the quality of what emerges from the sequence, never on movement through it.

Keeping the capability the method depends on. Sequencing requires people who can verify, and verification requires domain knowledge that is itself built by doing the work. An organisation that automates all of the underlying practice will eventually find it has nobody left who can staff a gate: the method assumes a capability that unreflective deployment destroys, which means the scaffolding needs maintenance as well as design.

The maintenance routine is the single most useful thing I do in my own work: periodically, and without assistance, reconstruct from scratch a proportion of the work you would normally generate with a model. Not as busywork, but as deliberate upkeep of a capability that passive supervision will otherwise allow to atrophy. For junior professionals this is construction rather than maintenance, and it needs designing into the role: interrogating competing outputs for logical consistency, tracing claims to primary sources, filtering synthetic volume. Those tasks require domain knowledge to perform and build it in the performing. Two conditions must hold: psychological safety in Edmondson's sense, so people can be visibly wrong without cost (Edmondson, 1999), and access to senior practitioners who can demonstrate the difference between careful and adequate work. Both are principally the work of middle managers, which is one more reason that eliminating them to pay for the technology is a poor trade.

Stated as rules, because a paper offering method should be willing to be prescriptive. Gate where errors become load-bearing and where context shifts weight. Iterate inside stages, never across the whole. Require generation, not approval, at every gate. Measure the output of the sequence, never the movement through it. And maintain, by deliberate unassisted practice, the capability the gates consume. None of these rules requires a new tool. All of them require a decision.

9. The Disciplines That Already Solved This

In my own practice I have consistently seen better output from people trained as engineers, scientists and formal logical thinkers, and I want to examine that observation rather than assert it, because a machine-learning practitioner praising the technically trained is marking his own homework. There is an obvious risk that I am recognising work that is legible to me rather than work that is better, and a second risk that technically trained

people apply these tools disproportionately to verifiable tasks, in which case I am observing a property of the task rather than the person.

Having sat with both objections, I think there is something real underneath, but the category is wrong. The relevant variable is not engineering. It is training in a discipline that makes its verification procedure explicit. Engineers decompose systems into subsystems with defined interfaces, which is methodological sequencing under another name; scientists are trained to treat a result as a distribution rather than a fact; both specify before building and iterate against the specification. They turn up pre-trained in the method this paper describes. But so, in their own idioms, do litigators with precedent and adversarial testing, doctors with differential diagnosis, auditors, historians with source criticism, and experienced editors. What these disciplines share is not a technical education. It is an external standard against which a claim can be run, and a professional habit of running it.

The interesting question is why some disciplines have that apparatus and others do not, and the answer is uncomfortable and useful in equal measure. The fields with explicit verification traditions built them because their failures were expensive and visible. Engineering failures fall down. Medical failures kill people. Audit failures get prosecuted. Those fields developed checklists, peer review, differential diagnosis, sign-off hierarchies and post-mortems because the alternative was catastrophe they could not conceal. Fields where failure is diffuse, slow to surface or difficult to attribute never had to build it. Much of management, strategy, marketing and internal advisory work runs on plausibility, because nobody could easily prove the advice wrong, and the feedback loop between a recommendation and its consequences is long enough that attribution rarely happens.

Then generative AI made plausibility worthless as a signal of competence, everywhere, at once.

That is why the fields with no verification tradition are the most exposed right now. They are exposed precisely because they never needed the immune system the engineering disciplines were forced to develop. It explains why the rubber-stamp review bites harder in professional services than in structural engineering, why workslop concentrates where it does, and why the historical anchors in these papers keep coming from aviation, cryptanalysis and spaceflight. I am not borrowing from those fields for colour. I am borrowing from them because they are the only ones that ever had to solve this.

The practical implication is not to hire engineers. It is to ask, for each significant category of work, what the verification procedure actually is, whether anyone can articulate it, and what would have to be true for a claim in that domain to be demonstrably wrong. Where there are answers, the AI method largely designs itself. Where there are not, the first task is not deploying a tool. It is building the standard that deployment will be measured against: scaffolding work of the oldest kind, erecting the external structure a trade never got around to building for itself.

10. Objections I Cannot Fully Answer

I would rather state these than have them stated for me.

The bitter lesson. The history of AI is fairly consistently a history of hand-crafted structure being outperformed by general methods with more compute (Sutton, 2019). If it holds here, everything in this paper is a transitional artefact. Part of my answer is in Section 7: the bitter lesson is a claim about which methods win in AI research, and it does not obviously extend to the return on personalised human method. The rest is that the scaffolding on the first path is an apparatus for verification and accountability rather than a workaround for capability. A better model lowers the error rate; it does not tell you who is answerable for the output, create the record a regulator will ask for, or supply context that exists only inside your firm. Those are not capability problems, so they do not scale away. Moderate confidence, and it would be more convincing with a decade of evidence behind it.

Agentic systems internalise the sequencing. Increasingly, systems decompose, plan, execute and review without a human orchestrating the loop. I think machine-run loops relocate the problem rather than solving it: a system that verifies its own work shares the failure mode of the work it is verifying, and the composite output still does not signal its own error. Arguably it signals less, since self-review adds another layer of fluent confidence. The gate criterion becomes more important in that environment, because the distance between the machine having done the work and somebody being answerable for it grows wider. But this is the objection most likely to age badly, and anyone building on this argument should watch it rather than assume I have settled it.

Gates cost something. If the value of the tool is speed and you insert six human checkpoints, you may have consumed the gain. There is no general answer; the right number of gates is a function of how expensive your errors are, and the honest way to run it is as an explicit trade: decide what an error costs in a given class of work, and buy only as much verification as that justifies. What I can report is that the organisations getting this wrong in my experience are almost never over-gated. Scaffolding sized for a cathedral should not be billed to a garden wall, but the reverse mistake is the one I actually meet.

This paper contains no measured example. The method itself is not hypothetical. I have spent three years implementing AI inside organisations across finance, IT, legal, sales, marketing, client delivery and operations, and the work has always reduced to the same three activities: training the people, choosing the right tools, and redesigning the workflows. The gated sequence in Section 8 is a working artefact of that practice, and in client scoping the first exercise is now routinely a verifiability map: rating candidate use cases by how checkable their outputs are, and pointing the tool only at the ones that pass. Drafting against a defined template rates as workable; cross-document consistency checking rates as weak; anything the machine would be trusted to pass or fail on its own is excluded outright. But what none of this yet has is numbers. The verification records show what the gates caught; they do not show what the sequence cost in additional time, or what would have shipped without it, because the work was not instrumented while it was being done. By this paper's own standard that gap matters. Assembling those measurements is the first obligation of the papers that follow, and until then the sequences here should be read as documented practice, not validated findings.

I will add the finding from those three years that I least expected. Very little of the hard part has been new. Every difficult problem in AI implementation has so far turned out to be

an old problem in change management wearing new clothes: cultural resistance, workflow design, expectation management. People first, then process, then tools, with the technology last. The scaffolding in this paper is my attempt to state what that old discipline means when the component being managed is probabilistic. If that is right, it is reassuring, because the discipline is well understood. It is also deflating, and I would rather admit the deflation than dress the method up as more novel than it is.

11. The Fundamentals, and the Organisation That Acts on Them

Four claims, with confidence attached. First: the output does not signal its own error, and this single property, rather than probabilistic behaviour in general, is what makes the technology organisationally difficult. High confidence. Second: human capability caps verification rather than generation, and for fluent-and-wrong output the economics of checking invert. High confidence in the mechanism; moderate confidence in how far it generalises. Third: reliability comes from gated sequence rather than better components, because that is how every other field solved the equivalent problem. High confidence in the principle; considerably lower in my own specific sequences, which are working methods rather than validated findings. Fourth: there are two paths requiring different structures. Moderate confidence, and it is the newest of the four in my own thinking, which is usually a reason for caution.

What would change my mind: sustained evidence that agentic systems can perform reliable self-verification on high-context work; evidence that model improvement closes the context-weighting gap in ways that survive contact with tacit organisational knowledge; or simply time, and the discovery that what I took for structural properties were features of a particular generation of tools. I would rather be caught having said this plainly than be found to have implied more certainty than I had.

If those fundamentals hold, the leadership task follows, and it is a question about people as much as processes. The banding of tasks proposed in the second paper, a small automated band, a large collaborative core, a protected human band, describes work. It needs a companion describing workers, because the same person moves between bands during a single day and needs a distinct posture for each. The four postures I see emerging owe an obvious debt to Dell'Acqua and colleagues, whose consultants split into Centaurs, dividing work with the machine at a clear boundary, and Cyborgs, integrating it throughout (Dell'Acqua et al., 2023); I have extended the taxonomy in both directions, and offer it as a working tool rather than a finished theory.

The Master of Machines designs the sequences and governs the gates. Their output is architecture rather than artefacts: the verifiability map, the placement of gates, the standards a gate artefact has to meet. They are the scarcest posture and the hardest to hire for, because the role requires both domain depth and the verification habits of Section 9, and the natural candidates are often mid-career practitioners whose job title gives no hint of it.

The Cyborg is natively integrated with the tools across most of their work, with oversight built into how they operate rather than bolted on afterwards. This is the second path made

flesh. The risk in this posture is drift, the slow slide from interrogating output to accepting it, which is why the maintenance routine of Section 8 applies to Cyborgs most of all.

The Copilot user engages tactically, reaching for assistance on bounded problems rather than as a default mode. Most of the workforce will spend most of its time here, and the organisation's obligation to them is not prompt training but clarity: a verifiability map that tells them, task by task, whether the machine may lead, may assist, or is excluded.

The Naked worker, a deliberately uncomfortable name, covers the tasks where the machine must not intrude at all, because the stakes are too high, the context too tacit, or the human relationship is itself the product. Naming this a posture converts what would otherwise be read as a failure to modernise into what it actually is: a designed decision, made in advance, about where the firm's judgment must remain visibly and accountably human.

Deciding which work, and which people, belong in which posture is the leadership task of the next five years, and it cannot be delegated to a technology function. It requires knowing what your organisation is actually good at, which is where this series began, and knowing what capability you are betting on retaining, which is where its second paper ended. The arc across the three is amplifier, rust and scaffold: the technology amplifies what you already have; unreflective deployment corrodes it; and method holds the two in a productive relationship rather than a destructive one.

I began on a runway in Ohio in 1935, with an aircraft that killed its pilot and a verdict that it was too much airplane for one man to fly. The response was not a smaller aircraft, and it was not a better pilot. It was an index card.

There is an objection buried in this paper's title, and I want to end by facing it, because it contains the most hopeful thing I have to say. Scaffolding comes down. That is the point of it: it is erected so that something can eventually stand without it. Some of what this paper prescribes will come down too. Gates will consolidate as the models improve, sequences will simplify, parts of the method will migrate into the tools themselves, and anyone who tells you their framework is permanent is selling you the framework. But consider what scaffolding leaves behind when it comes down properly. Masons worked for years at heights that would have killed them, on timber that appears in no painting of the finished cathedral, and the vault stands because the structure that built it was taken seriously by everyone who climbed it. What remains afterwards is not the absence of the scaffold. It is a building, and a workforce that knows how to raise the next one.

The engine is here, it is extraordinary, and it is too much airplane to fly on instinct. Build the scaffolding. Work at the height it makes possible. And when parts of it come down, let that be because your organisation can finally hold the shape on its own, and not because nobody ever put it up.

A Note on Sources

Several figures in this paper come from research that is recent, preliminary, or both. The MIT NANDA study has not been peer reviewed and its methodology has been publicly contested; I use its headline figure as one signal among several, and the first paper in this

series discusses its limitations at greater length. The Ramp and Revelio employment analysis covers a twenty-four-month window and a population skewed toward fast-growing, venture-backed firms. The workslop cost figures are self-reported. The verification-inversion argument in Section 6 is presented as synthesis of a convergent literature rather than as an original finding; the citations there identify the closest published statements I am aware of, and I would welcome earlier ones. The claim that AI widens performance on unverifiable work is a prediction consistent with, but not yet demonstrated by, the published evidence, and I have marked it as such. One clarification for readers who cross into the technical machine-learning literature: there, "verification asymmetry" usually names the opposite phenomenon, models checking candidate answers more cheaply than they generate them, which is the basis of test-time scaling (Venktesh, Rathee & Anand, 2025). The two claims are compatible: the same technology that makes bounded answers cheap for machines to check is making unbounded claims expensive for humans to check. The client practice described in Sections 8 and 10 is drawn from live delivery work and has been anonymised; no client is named and none should be inferred. It is described rather than measured, and Section 10 addresses why that matters. The gate placement rules, the maintenance routine, the four postures, and the observation about verification traditions in Section 9 are one practitioner's working method, offered for testing rather than deference.

References

Acemoglu, Daron. "The Simple Macroeconomics of AI." Economic Policy 40, no. 121 (2025).

An, Tao. "AI as Equalizer or Amplifier? Task Complexity as the Moderating Factor for Human Expertise in Hybrid Intelligence Systems." arXiv:2512.10961, December 2025.

Bainbridge, Lisanne. "Ironies of Automation." Automatica 19, no. 6 (1983): 775–779.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. "Generative AI at Work." National Bureau of Economic Research Working Paper 31161, 2023.

Challapally, Aditya, Chris Pease, Ramesh Raskar, and Pradyumna Chari. The GenAI Divide: State of AI in Business 2025. MIT Project NANDA, July 2025.

David, Paul A. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." American Economic Review 80, no. 2 (1990): 355–361.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013, 2023; published in Organization Science, 2025.

Edmondson, Amy C. "Psychological Safety and Learning Behavior in Work Teams." Administrative Science Quarterly 44, no. 2 (1999): 350–383.

European Parliament and Council. Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act). Official Journal of the European Union, 2024.

Garousi, Vahid. “Human Oversight and Overload: Two Hidden and Costly Burdens of AI-Assisted Software Engineering.” arXiv:2606.05770, 2026.

Gawande, Atul. The Checklist Manifesto: How to Get Things Right. Metropolitan Books, 2009.

Graphite. AI Now Writes as Many Online Articles as Humans. May 2026.

Huang, Lingxiao, Wenyang Xiao, and Nisheeth K. Vishnoi. “Delegation and Verification under AI.” arXiv:2603.02961, 2026.

Jevons, William Stanley. The Coal Question. Macmillan and Company, 1865.

Kahneman, Daniel. Thinking, Fast and Slow. Farrar, Straus and Giroux, 2011.

Kharazian, Ara, Lisa Simon, and Ryan Stevens. “A New Look at AI’s Impact on Jobs: Firm-Level AI Spending and Workforce Adjustment.” Ramp Economics Lab Working Paper, 30 June 2026.

Ma, Jiaqi W. “Toward an Engineering of Science: Rebalancing Generation and Verification in the Age of AI.” arXiv:2605.10425, 2026.

MIT Sloan School of Management. “Seeing Real Value from AI Depends on Being Able to Verify Its Outputs.” Ideas Made to Matter, 2026.

Niederhoffer, Kate, Gabriella Rosen Kellerman, Angela Lee, Alex Liebscher, Kristina Rapuano, and Jeffrey T. Hancock. “AI-Generated ‘Workslop’ Is Destroying Productivity.” Harvard Business Review, 22 September 2025.

Perez, Carlota. Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages. Edward Elgar, 2002.

Polanyi, Michael. The Tacit Dimension. Doubleday, 1966.

PwC. 2026 Global AI Jobs Barometer. PricewaterhouseCoopers, June 2026.

Reason, James. Managing the Risks of Organizational Accidents. Ashgate, 1997.

Sutton, Richard S. “The Bitter Lesson.” Incomplete Ideas (blog), 13 March 2019.

Venktesh, V., Mandeep Rathee, and Avishek Anand. “Trust but Verify! A Survey on Verification Design for Test-time Scaling.” arXiv:2508.16665, 2025.

von Neumann, John. “Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components.” In Automata Studies, edited by Claude E. Shannon and John McCarthy, 43–98. Princeton University Press, 1956.