Cognitive Rusting

What Organisations Lose When They Automate Their People

Conor Flynn | Trinzo | July 2026

Executive Summary

Most organisations are now running two AI transformations at once. The official one lives in board papers: a roadmap, a governance committee, a pilot programme, a slide with a curve on it. The unofficial one lives at desks, where capable people are quietly using consumer AI tools on their own accounts, getting real work done with them, and hiding the fact from their managers. The contrast between what is said in boardrooms and what is happening at desks is stark, and the gap between the two is where this paper begins.

The paper is about what the official transformation, deployed unreflectively, does to the human capability of an organisation over time. I call it cognitive rusting: a slow, invisible degradation of the skills, the knowledge and the professional judgment that a firm took decades to build. The argument, reduced to a sentence, is that organisations are automating the very work through which professional judgment gets built, and the capability they are quietly dismantling is the same capability that determines whether any of this works at all.

Four claims carry the argument. The first is that the most common human-in-the-loop workflow, in which the machine generates and the human reviews, is not a safeguard. It is a trap, and there is forty years of evidence behind that sentence. The second is that the entry-level and middle layers of the organisation are load-bearing in ways the efficiency case does not capture: the junior roles being eliminated were the apprenticeship through which judgment was built, and their disappearance carries a commercial cost that never appears on a productivity dashboard. The third is that the case against middle management misunderstands what middle managers actually do; their real work cannot be automated, only abandoned. The fourth is that all of this is happening while organisations bind their core processes to a small number of AI systems they do not control, cannot inspect, and could lose access to with limited notice, so that the deskilling and the dependency compound each other into a category of risk that most operational frameworks have not yet named.

This paper is part of a series of three, but it stands alone. Where an argument continues elsewhere, I will say so briefly and move on.

Introduction

I have spent fifteen years working in machine learning. Long enough to have seen the cycle before, in a smaller and quieter key: the excitement of a new capability, the rush to apply it,

the gradual discovery that the hard part was never the technology. It was always the people, the processes, and the institutional conditions that determined whether the technology delivered or disappointed.

The current moment is different in scale. The pressure to deploy is more intense, the capital at stake is larger, and the claims being made are more ambitious. But the underlying dynamic is familiar, which is why this paper keeps reaching for historical parallels, from paint tubes to power looms to personal computers. History never repeats, but it hums the same tune.

This paper is written for senior leaders who are responsible for the financial and operational health of their organisations over the medium and long term. It is not a technology paper, and it deliberately does not contain an implementation methodology. It is an attempt to describe clearly what is happening inside organisations that are deploying AI quickly and unreflectively, and what they will wish they had known five years from now.

Some of what follows rests on published research, and I will cite it. Some rests on what I see inside client organisations week to week, which is vivid but unsystematic. Where I am drawing on practice, I will say so. Where I am uncertain, I will say that too.

Section 1. The Gap Between the Boardroom and the Desk

The transformation programme, as presented, is impressive. A roadmap with phases. A secure, approved platform. A centre of excellence. Meanwhile, three floors down, an analyst is pasting client data into a free chatbot on her phone, because the approved platform does not do what she needs and the request form for new tools takes six weeks. She is not an outlier or a laggard. In my experience she is one of the most capable people on her team, which is exactly why she found the workaround first.

I see this in nearly every organisation I work with. I cannot give you a reliable percentage, and I distrust the surveys that claim to, because people do not answer honestly about behaviour they are concealing from their employer. What I can tell you from practice is that when I sit with frontline teams and the managers leave the room, the stories change. People describe real productivity gains from tools they are not supposed to be using, and they describe hiding those gains, working two ways at once: the official way for the audit trail, the unofficial way for the actual work.

Why hide something that works? The answer I hear is a mixture of shame and pragmatism. Shame, because using the tool can feel like cheating, and because admitting how much of the job it can do feels dangerous in a year of headcount reviews. Pragmatism, because trying something unsanctioned invites scrutiny, and scrutiny lands on the individual rather than on the policy that made the workaround necessary. These are people carrying rising workloads with static resources. They are drowning, and reaching for whatever is nearby.

We have been here before. In the early 1980s, middle managers smuggled personal computers into offices on their own credit cards, because the official route to computing power ran through an IT queue measured in years. The organisations that responded by locking down procurement drove the machines under desks and into cupboards. The ones

that listened to what their most motivated people were telling them got a decade of advantage for it. Tighter surveillance does not stop unsanctioned use. It moves the behaviour somewhere less visible and less safe, which is the worst available outcome for a firm that handles client data.

The shadow AI economy matters here because of what it reveals. The official transformation is being designed around a workflow that quietly corrodes the people inside it, while the unofficial one shows what active, engaged use of these tools actually looks like. The rest of this paper is about the first problem, and about why the instinct behind the second is the thing worth building on.

Section 2. The Reviewer Trap

The most common human-and-machine workflow design looks, on first inspection, like a sensible division of labour. The machine generates the draft. The human reviews it. The output goes forward. This arrangement is widely described as responsible AI deployment. It is nothing of the kind. It is a trap, and it is the single most consequential design error I encounter.

The problem is not that the human reviews the machine output. The problem is what the human actually does when they review it, and what that role does to them over time.

Forty years ago, the British researcher Lisanne Bainbridge wrote a paper called Ironies of Automation (Bainbridge, 1983). She was studying industrial control systems, chemical plants and aircraft cockpits, but her findings apply with uncomfortable precision to the office of 2026. Bainbridge identified two structural problems with automation designs that retain a human in a monitoring role.

The first is the skill atrophy irony. You automate a process because you believe the machine is more reliable than the human. But you keep the human in the loop to intervene when the machine fails. The automation then deprives the human of the very practice they would need to intervene effectively. When the machine eventually encounters something it cannot handle, the human's skills have atrophied to the point where they cannot step in. The safety net is there on paper. It is not there in practice.

The second is the vigilance irony. If a system rarely fails, human attention drifts. This is not a character flaw. It is biology. We are not built for sustained, effortful attention to systems that mostly behave as expected. The more reliable the AI output becomes, the less critically the reviewer examines it, and the errors that get through are not the obvious ones. They are the subtle ones, the contextual misalignments and quiet logical faults that require real engagement to catch, precisely the kind of engagement the reviewer role systematically discourages.

Daniel Kahneman's distinction between System 1 and System 2 thinking explains the mechanism (Kahneman, 2011). System 1 is fast, automatic and effortless. System 2 is slow, deliberate and exhausting. Generative AI produces fluent, authoritative, grammatically confident text, and that surface confidence triggers System 1 in the reviewer. Critical review requires System 2. Under deadline pressure and cognitive fatigue, System 1 wins.

Here is what the role actually looks like from inside. You skim, you tidy the formatting, you soften a phrase, you move it along. It feels like diligence and it looks like diligence. The substantive decisions, the structure of the argument, the selection of evidence, the framing of conclusions, were made by the machine, and the human approved them on the basis of plausibility rather than verification. It is not laziness. It is what the role quietly trains people to do. When I ask teams how often they truly review a colleague's work as thoroughly as the approval chain suggests, the question draws a nervous chuckle from the team and an outright laugh from leadership. I am guilty as charged myself. The honest test of any review step is whether the professional is still in the work, or just near it.

This matters in a knowledge economy in ways it did not in an industrial one. A factory worker reduced to monitoring a machine is disengaged, which is unfortunate, but the machine still produces its output. A professional reduced to reviewing output they did not shape and cannot meaningfully interrogate is being stripped of the thing that makes them professionally valuable: the active exercise of expert judgment. In knowledge work, the human mind is not the operator of the process. It is the process. Reduce it to passive supervision and you have not made knowledge work more efficient. You have made it look like knowledge work while quietly replacing its substance. The role becomes, in effect, a coal mine for the mind: extractive, repetitive, and corrosive to the very capacity it depends on.

There is a legal dimension that senior leaders should take personally. Madeleine Clare Elish described what she called the moral crumple zone: the human positioned nominally in control of an automated system, who absorbs the blame when the system fails (Elish, 2019). Just as the crumple zone of a car deforms to protect the passengers, the passive reviewer deforms to protect the organisation from liability when the AI output causes harm. They did not produce the output. They could not meaningfully have verified it. But they approved it, and that approval is the legal fig leaf behind which the organisation shelters. Nassim Taleb's principle of skin in the game sharpens the point: robust systems require that the people who make decisions share in their consequences (Taleb, 2018). The passive reviewer model severs that link entirely.

Regulation is now catching up with this, faster than most organisations realise. The EU AI Act (Regulation 2024/1689) entered into force in August 2024 and is being phased in: prohibitions and AI literacy obligations from February 2025, general-purpose model obligations from August 2025, and the bulk of the high-risk system requirements from 2 August 2026, weeks after this paper's publication. Article 14 requires that high-risk systems be designed so that humans can effectively oversee them, and it names the failure mode directly: overseers must remain aware of the tendency to automatically rely or over-rely on the system's output, which the regulation calls automation bias. The drafters, in other words, had read their Bainbridge. The high-risk category is defined by use case rather than by sector, and it includes credit scoring, risk assessment and pricing in life and health insurance, recruitment and employment decisions, and emergency triage, along with AI used to assist courts. A passive reviewer who rubber-stamps machine output does not satisfy the plain intent of that requirement, and for organisations deploying AI in those contexts this stopped being a future problem on the day the obligations took effect.

The alternative is not to remove the human. It is to redesign where and how the human engages: as the generator of the core assumptions, the interrogator of the machine's output, and the accountable decision-maker about what goes forward. Shoshana Zuboff saw this fork in the road in 1988, studying how organisations absorbed information technology (Zuboff, 1988). She called it the choice to automate or to informate. To automate is to use technology to reduce human agency and cut labour cost. To informate is to use technology to make work more visible and more intelligible to the people doing it, so that their judgment deepens rather than decays. The reviewer trap is the logical endpoint of the automate path. Everything this paper recommends is, one way or another, a decision to informate instead.

Section 3. The Cobra Effect

The story is usually told like this. During British colonial rule in India, the administration in Delhi grew concerned about venomous cobras and offered a cash bounty for every dead one. Cobra heads arrived in volume. Then citizens discovered a more efficient business model: breeding cobras at home for the bounty. When the administration cancelled the programme, the breeders released their stock, and the city ended up with more cobras than it started with.

I should be honest about this story, in a paper that argues for verification. There is no primary source for the Delhi bounty. It entered wide circulation through the economist Horst Siebert, who used it as the title of a book on perverse incentives (Siebert, 2001), and it may well be apocryphal. I tell it anyway, clearly labelled, because it is the most vivid available illustration of a mechanism that is thoroughly documented: when a measure becomes the target, people optimise the measure rather than the purpose behind it, an observation formalised by Charles Goodhart in the context of monetary policy (Goodhart, 1975) and confirmed many times since. The parable may be invented. The mechanism is not.

Organisations deploying AI under pressure to demonstrate productivity gains are breeding digital cobras at considerable scale.

A software team evaluated on code shipped will use agents to generate large volumes of syntactically plausible code that passes surface testing while accumulating architectural debt and security vulnerabilities invisible to any reviewer not deeply familiar with the system. A customer service team measured on ticket closure will deploy automated responses that close tickets without resolving the underlying issue, producing metrics that look excellent and satisfaction scores that will deteriorate a quarter later. A consulting team incentivised for rapid, comprehensive deliverables will assemble fifty-page briefs in hours, fluent and confident and filled with assertions a knowledgeable client will see through immediately. In each case the metric is being optimised while the purpose behind it erodes, and because the erosion is gradual and the metrics look fine, the problem stays invisible until something forces it into view: a client complaint, a security incident, a regulatory examination, or the retirement of the senior people who knew what good work looked like.

The corrective, when it comes, tends to be painful. The organisation restricts the tools that generated the problematic output, and the cobras are released: a large inventory of synthetic artefacts that nobody fully understands, produced by processes that left no institutional trace, validated by people who were in no position to validate them.

The antidote is not better surveillance or cleverer metric design, though both help at the margin. It is designing workflows in which the human's intellectual contribution is actually required, visible, and recorded. When the human cannot be removed from the substantive process without the output degrading, the incentive to game the metric collapses. There is nothing to game.

Section 4. Why Firms Exist

Ronald Coase asked in 1937 why firms exist at all, when markets could in principle coordinate everything (Coase, 1937). His answer was transaction costs: organising work inside a hierarchy is cheaper than negotiating every task across an open market. The firm exists because internal coordination is cheaper than market coordination.

Agentic AI systems that draft, route, schedule and execute workflows drive the cost of explicit coordination towards zero. If a middle manager's function were primarily to relay instructions, collate status updates and move information between layers, the economic rationale for that function really has weakened, and Coase's own framework says so. The mistake, and it is the central mistake of the flatten-the-organisation argument, is to treat explicit coordination as the whole of what human intermediation does. Explicit coordination is codifiable and machine-amenable. Tacit synthesis, the work of translating strategy into operational reality, managing relationships, and carrying the institutional memory of the firm, is neither. Automating the first does not make the second unnecessary. It makes it more important, because it strips away the routine work that used to hide it.

There is a second economic effect worth stating plainly, and it has a long pedigree. William Stanley Jevons observed in 1865 that more efficient steam engines did not reduce Britain's coal consumption; they increased it, because making steam power cheaper made everyone want more of it (Jevons, 1865). The same dynamic is now running through cognitive work. When generating a document costs almost nothing, organisations do not produce the same number of documents more cheaply. They produce vastly more documents. Generation has become cheap; judgment has not; and so the bottleneck in the modern organisation has moved from the capacity to produce content to the human capacity to read it, weigh it, and act on what it contains. The practical response is not to generate less but to filter better, and filtering at that scale must itself be partly automated. The human's role shifts from reviewer of everything to architect of the filtering system, which is a more demanding job, not a lesser one, and it demands more investment in human capability, not less.

The firm that automates its coordination has not answered Coase's question. It has sharpened it, and the honest answer now depends entirely on the humans it kept.

Section 5. Cognitive Rusting and the Talent Pipeline

In the nineteenth century, a painter's apprentice spent years grinding pigment by hand. Then, in the 1840s, ready-made paint in metal tubes arrived, and the grinding stopped. The grinding was never the point; the deep familiarity with colour, texture and material was. And yet the craft did not collapse. Apprentices did not stop painting. They started painting sooner, at a higher level of abstraction, and the baseline of the profession shifted upward. Art historians credit the tube with helping to make Impressionism possible: painters could at last work quickly, outdoors, in front of the subject.

That is the optimistic case for automating the bottom rung, and I want to state it at full strength because it is often true. Technology removes drudgery, the apprenticeship reforms around a higher starting point, and the profession gains. It is not, however, guaranteed. When the power loom arrived, the skilled hand weavers were not elevated to a higher abstraction. They were bypassed entirely, and their accumulated expertise was economically stranded. When spreadsheet software reached corporate offices in the early 1980s, the finance profession expanded, but many of the ledger clerks whose work was automated were never offered the bridge to the new roles. They did not move up the value chain. They left. The difference between the paint tube and the power loom is not the technology. It is whether anyone redesigned the apprenticeship on purpose.

Which brings me to what is happening now. The entry-level tasks that AI is most readily replacing, the summarising, the first drafting, the data formatting, the document review, are precisely the tasks through which junior professionals built pattern recognition and domain intuition. The educational theorists Jean Lave and Etienne Wenger called this legitimate peripheral participation: novices become experts by doing real but low-stakes parts of real work, under the eyes of people who know the difference between good and not good (Lave and Wenger, 1991). Juniors do not learn judgment from a curriculum. They learn what good actually looks like by repeatedly doing work that is not yet good.

For the first time, there is large-scale evidence that this rung of the ladder is actually being removed rather than merely threatened. Researchers at the Stanford Digital Economy Lab, working with payroll records from the largest payroll software provider in the United States, found that since the widespread adoption of generative AI, early-career workers aged 22 to 25 in the occupations most exposed to AI have experienced a relative employment decline of around 16 percent, while older workers in the same occupations, and young workers in less exposed occupations, held steady or continued to grow (Brynjolfsson, Chandar and Chen, 2025). Two details of that study deserve attention. The decline shows up in employment rather than in wages, which is what you would expect if firms are simply not hiring at the bottom rather than repricing the work. And it is concentrated in occupations where AI automates the work rather than augmenting it, which is precisely the automate-or-informate fork that Zuboff identified four decades ago, now visible in payroll data.

I want to handle this evidence the way I am asking organisations to handle their own. The finding is contested, as any early finding about a fast-moving phenomenon should be. The authors themselves caution against reading AI as the sole cause; in a follow-up analysis

they showed that under their most stringent statistical controls the divergence becomes clearly significant only from 2024, and other researchers using different data and methods find smaller effects or different timing. The honest summary is that the canary data is consistent with the mechanism this paper describes, not proof of it. But an executive waiting for proof should notice what is visible on the ground: in the United Kingdom, the large accounting firms were reported through 2024 and 2025 to be cutting their graduate intakes, Deloitte by 18 percent and EY by 11 percent, while one job platform recorded entry-level vacancies in finance falling by roughly half, with AI cited as the principal cause (Myers, 2025). The direction of travel is not in serious dispute. What is in dispute is only how fast, and how much of it is AI.

So here is the claim, stated as plainly as I can make it. The entry-level and middle layers of the organisation are load-bearing in ways the efficiency case does not capture. Removing them does not produce a leaner version of the same firm. It produces a firm that has cancelled its own apprenticeship while the dashboard still looks healthy.

In the professional services and financial firms I know best, the graduate intake, the milk run as the industry calls it, is not primarily a source of cheap labour, and the firms that run it well have always known something that the efficiency case misses entirely. Most of those graduates will leave within a few years, and that is not a flaw in the programme. The ones who leave move into the organisations that later hire professional services firms, and a good number come back as clients. The graduate programme is not only a talent pipeline. It is a client-development pipeline. This mechanism is documented beyond my own observation: research on corporate alumni programmes finds that well-managed departures reliably convert former employees into customers, referrers, boomerang hires and ambassadors, which is why the most sophisticated firms invest in offboarding almost as carefully as onboarding (Dachner and Makarius, 2021). The consulting industry's most celebrated alumni network has functioned as a client-generation engine for decades, and every partner in a large firm can confirm the mechanism from their own contact list. A firm that hollows out its intake to capture an AI efficiency gain is therefore not just weakening its future capability; it is quietly shrinking its own future client base. For a chief financial officer, this second-order commercial cost may be the strongest argument in this paper, and it will never appear on a productivity dashboard.

The people running these firms know it, too, and the more thoughtful among them are worried. At a conference last year, a vice president at one of the Big Four told me plainly that he is afraid for his firm's business model within five years if they do not keep taking in graduates, training them properly, and building a place those graduates actually want to stay. What struck me was not the fear but the difficulty of acting on it: nothing is on fire yet, the intake reductions all look individually sensible, and the damage will surface on someone else's watch. It was a conversation between talks, not a study, and I report it as such. But set it beside the payroll data and the intake numbers, and it reads less like one man's anxiety and more like an early damage report. That is what cognitive rusting looks like from the inside. Not a crisis. A series of reasonable decisions.

Now the counterargument, because it is a serious one and it deserves room. Skills have always abstracted upward. Socrates worried that writing would destroy memory, a

concern we know, with some irony, only because Plato wrote it down (Plato, Phaedrus). Senior architects worried that CAD software would produce professionals who could not understand spatial relationships. In both cases the cognitive baseline shifted upward rather than collapsing. Today's junior architect cannot hand-draft a blueprint and does not need to; they work at a higher level of abstraction and produce more complex buildings. Perhaps the AI-era junior professional will likewise skip the drudgery and start at synthesis, exactly as the painter's apprentice skipped the grinding.

I find this reassuring in the long run, and I want to be honest that it may simply be correct. My answer is not that the abstraction will fail. It is that the abstraction is currently not being managed, and the paint-tube outcome was never automatic. The architects who worried about CAD were mostly wrong, but the firms that thought carefully about what junior architects should be learning in the CAD environment produced better architects than the firms that thought about it not at all. The question that matters is what the new cognitive apprenticeship should contain, and almost nobody deploying AI at scale is asking it.

Matt Beane's research shows what happens when nobody asks it. As robotic surgical systems concentrated the work of an operation in the hands of a lead surgeon at a console, residents who had learned by physically assisting were relegated to watching (Beane, 2019). The expert and the novice were decoupled. And the residents who still became skilled did so through what Beane calls shadow learning: unsanctioned practice at the margins, on simulators not built for the purpose, in procedures they were not supervised to perform. The learning did not stop. It went underground, and it became less safe. His broader research argues this pattern is now general across AI-exposed professions (Beane, 2024). Note the echo of Section 1: when the official system blocks what capable people need, they do not stop. They disappear from view.

From my own practice, three conditions look essential for any organisation redesigning its entry-level function around AI. Juniors must be trained to interrogate AI output rather than accept it. They must hold visible roles in the verification and governance of AI-assisted work. And they must remain in direct working contact with seniors who can show them what good looks like and why. I hold these with high confidence as a mechanism and moderate confidence in how far they generalise across sectors, and I will name my own bias before a reader does: I run a consultancy whose service is judgment-preserving AI adoption, so I am professionally disposed to find this problem everywhere. The reader should weigh that. I have weighed it, and I have watched too many graduate programmes quietly shrink this year to conclude that I am seeing only what I am paid to see.

Section 6. The Middle Management Core

A strand of thinking in the AI conversation treats middle management as a problem to be solved. The argument runs: middle managers route information, coordinate tasks and relay decisions; AI does those things more efficiently; therefore the layer is redundant, and the organisations that flatten first win. Section 4 explained why this argument is half right, which is what makes it dangerous. The half it gets right, explicit coordination, is real and automatable. The half it misses is the actual job.

Middle managers are the only people in an organisation with simultaneous access to the strategic intent of senior leadership and the operational reality of frontline work. They do not merely relay between those two domains. They translate. The instruction that arrives from above as an abstract priority has to be converted into something operationally meaningful for a team with specific clients, systems and constraints, and the account of frontline reality that travels upward has to be converted into something leadership can act on. That conversion requires contextual judgment that no current system can replicate. Quy Huy's research found that middle managers are the most vital allies of radical change precisely because they hold the informal networks, operational knowledge and relationships that neither executives nor frontline workers possess independently (Huy, 2001). Two decades later, with the gap between what leadership imagines AI can do and what frontline teams can actually do with it as wide as it currently is, that translation function is worth more, not less.

Ikujiro Nonaka and Hirotaka Takeuchi described organisational knowledge creation as a continuous cycle in which tacit knowledge is socialised, articulated into explicit form, combined, and internalised back into practice (Nonaka and Takeuchi, 1995). Automated systems are excellent in exactly one phase of that cycle, the combination of explicit, structured information. They are essentially absent from the rest. The work of absorbing qualitative frontline experience, converting it into structured insight, and translating it back into guidance people can use requires human presence at every stage, and middle managers are its natural home. Organisations that eliminate them do not eliminate the work. They leave it undone, or they redistribute it: downward onto frontline staff who lack the position to do it, upward onto executives who lack the time and who now also lack the operational visibility the middle layer used to provide. The firm does not become flatter and faster. It becomes more fragile and less honest with itself.

The middle layer is also where the rest of this paper converges. Middle managers decide, workflow by workflow, whether the humans in their teams are in the work or merely near it; they are the seniors whose attention makes the apprenticeship of Section 5 function; and they are the people best placed to notice digital cobras while the dashboard still looks healthy. The practical implication: stop evaluating middle managers on the parts of the role AI can do, the routing and the status tracking, and start evaluating and developing them on the parts it cannot: translation, mentorship, institutional memory, and steadying a workforce under technological pressure. That is a larger change to performance management than it sounds, and I have yet to see an organisation complete it. The first ones that do will have built the connective tissue an AI-era firm actually runs on.

Section 7. What the Firm Is Actually For

If AI can coordinate tasks at near-zero cost, draft at industrial scale, and increasingly filter its own output, the question deserves to be asked directly: what does the firm still provide that justifies its existence?

Michael Polanyi's answer, given decades before the question, was tacit knowledge: the knowledge embedded in practice and held in the judgment of skilled people, which cannot be fully written down (Polanyi, 1966). You can read every book on bicycle riding and still

fall off. The knowing is in the riding, and it transfers only through practice. Professional expertise is largely of this kind. The experienced lawyer who reads a contract and immediately senses where the risk sits cannot always articulate why. The consultant five years into a client relationship knows things about its culture that no briefing document contains. The manager three years into a team knows exactly who will hold in a crisis. That knowledge is real, it is valuable, and it is precisely what a model trained on the written record does not contain, because it was never written down.

In an economy where fluent, plausible text costs nothing, the premium on this kind of knowledge rises. Clients do not engage professional firms to receive AI-generated analysis, however good; they can generate that themselves. They engage firms for the accountable, experientially grounded human judgment that stands behind the analysis and takes responsibility for it, with something to lose. This is the deeper answer to Coase. The firm survives not as a coordination mechanism, which AI can replicate, but as a trust mechanism, which it cannot. Trust needs a person who understood the work, staked their name on it, and can be held to account; an approval click from someone who skimmed the output provides the appearance of that and none of the substance. An organisation can automate its way out of coordination costs. If it automates its way out of the capacity to underwrite its own work, it has automated the reason clients pay for it, and what remains is something very efficient and very fragile.

Section 8. The Dependency Question

Everything argued so far concerns what automation does to people. This section concerns what dependence does to the organisation, because the two exposures are compounding each other and almost nobody is pricing the combination.

Consider what the modern AI supply chain actually looks like from a desk in Dublin, or Frankfurt, or Singapore. The frontier models that matter are built by a handful of firms, most of them American, one or two Chinese. They run on accelerator chips designed substantially by a single company, fabricated overwhelmingly by a single manufacturer in Taiwan. I take no view here on how cross-strait relations resolve; the point is narrower. A concentration like that, found in any other supply chain, would sit at the top of the risk register on its own. Access to all of it is mediated by commercial terms that can change at contract renewal, and by export and trade policy that can change overnight. A professional firm that has wired a frontier model into its core workflows has, whether it thinks of it this way or not, made its ability to serve clients contingent on the foreign policy of other states and the strategic choices of a few private companies it has no relationship with beyond an API key.

If that sounds abstract, consider how the access regime itself has behaved. The United States began restricting exports of advanced AI chips in 2022 and has tightened, loosened and redrawn those restrictions repeatedly since. In January 2025 it published a Framework for Artificial Intelligence Diffusion that sorted the world's countries into tiers with different rights of access to American AI technology; four months later the framework was rescinded before it ever took effect, with a replacement promised (Bureau of Industry and Security, 2025a, 2025b). I take no position here on the merits of either move. The point is the

metabolism. The rules governing who may use the most capable AI systems, and where, changed twice within a single year, and there is no reason to believe they have stopped changing. An organisation building twenty-year client relationships on top of that regime is building on ground that moves.

Nor does the disruption require geopolitics. Models are deprecated on the provider's schedule, not the customer's, and a replacement model is not the same model: workflows validated against one system's behaviour can degrade quietly when the system underneath them changes, which is a compliance problem as much as a quality one. Pricing and usage terms shift. And the infrastructure itself fails. In July 2024 a single faulty software update from one security vendor disabled roughly eight and a half million machines worldwide in a morning, grounding flights, cancelling surgeries and stopping broadcasters mid-transmission. That incident had nothing to do with AI, and everything to do with what happens when a great many organisations depend on one supplier and have no rehearsed way of operating without it. It is worth sitting with the thought of an equivalent outage, or an equivalent withdrawal, striking a model that a firm's entire drafting, analysis and client-response capability now silently routes through.

The financial regulators, to their credit, saw this shape coming. The EU's Digital Operational Resilience Act, which has applied to financial entities since January 2025, exists precisely because regulators concluded that the sector's dependence on a small number of critical technology providers had become a systemic risk; it obliges firms to map their critical third-party dependencies, assess concentration risk, and maintain tested exit strategies (Regulation 2022/2554). Very few firms I encounter have classified their AI model providers as critical third parties under that lens. On any honest reading of how these tools are now used, they are becoming exactly that, and I would expect supervisory attention to arrive at that conclusion before most risk registers do.

Here is where the two halves of this paper meet, and why I have given the dependency question a section of its own. Dependency alone is a procurement problem, and procurement has known answers: multiple providers, open-weight fallback models, abstraction layers that let you swap one system for another, contractual notice periods. Deskilling alone is a capability problem, and this paper has been about its answers. But dependency multiplied by deskilling is a different animal. An organisation whose core workflows require a specific external model, and whose people can no longer perform the underlying tasks without it, has no fallback at any price. The outage, the export restriction, the deprecation, the tenfold price increase arrives, and the firm discovers that the human capability it optimised away was not a redundancy. It was the business continuity plan. Its absence converts a supplier problem into an existential one.

The questions a board should be asking are not exotic. Which of our revenue-critical workflows depend on a single external model, and what is our tested degraded mode: a second provider, an open-weight model run on our own infrastructure, or humans doing the task more slowly? How long would it take to revalidate our work products on a substitute system, and who signs off that the substitution is safe? And the question underneath the others: who in this organisation still knows how to do the task at all, and what are we doing to keep that true? For most organisations I meet, the honest answers

are: we have not mapped it, we do not know, and nobody has owned the question. Business continuity planning in this area is still asking about data centres and backup power. The dependency that matters now is cognitive, and it is being deepened by exactly the deployment pattern this paper describes, one efficiency gain at a time.

Section 9. Building Differently

If the diagnosis is roughly right, what should organisations actually do? I offer five orientations rather than a framework, because frameworks imposed from the top without local understanding are how you breed cobras.

The first is to accept that AI use will be idiosyncratic. These systems respond differently to different users, prompts and contexts, and most of your staff are already using them, sanctioned or not. Standardising use through rigid prompt libraries and monitoring individual usage is operationally counterproductive and psychologically corrosive, and Section 1 showed where it leads: not to compliance but to concealment. Creating safe, sanctioned space for experimentation, and treating your shadow AI users as scouts rather than offenders, produces better outcomes than prohibition ever will.

The second is to design for active engagement rather than passive review. Every AI-involved workflow should face one question: is the human an active contributor to the substance of the work, or a reviewer of output they did not shape? If the latter, redesign it. The banding I use in practice divides AI-assisted work into three tiers. Roughly ten percent can be fully automated: tasks where errors are systematically catchable and individually cheap. Roughly eighty percent should be genuine collaboration, where the human sets the frame, interrogates the output, challenges the assumptions and makes the substantive decisions. The remaining ten percent, where stakes, regulation or ethics demand it, stays fully human-authored. The exact proportions vary by function and sector, and I hold the numbers loosely; the principle, that the human must be in the work rather than nominally above it, I do not hold loosely at all.

The third is to treat the talent pipeline as a capital investment. Before eliminating an entry-level function, the question is not whether AI can perform the task; it almost certainly can. The questions are what capability was being built in the people who performed it, where that capability will now come from, and, for the reasons Section 5 gave, what the intake was quietly earning the firm commercially in future clients and future leaders. If those questions have no good answers, the saving is being taken out of capital, not cost.

The fourth is to invest in middle management specifically: not the coordination functions AI absorbs, but the translation, mentorship and institutional memory that it cannot. This means changing what middle managers are evaluated on, which is a deeper intervention than a training programme and worth many times more.

The fifth is to build for resilience rather than only efficiency, and to treat Section 8 as an operational agenda rather than a geopolitical aside. Map the model dependencies in revenue-critical workflows. Establish and actually test a degraded mode, whether that is a second provider, an open-weight fallback, or human performance at reduced volume. Put AI providers through the same critical-third-party lens that regulation already applies to

other technology dependencies. And recognise that the cheapest and most durable form of redundancy is the one this whole paper has been defending: people who still know how to do the work. The best positioned organisations in three to five years will not be the ones that removed the most human labour. They will be the ones that kept real human capability alongside the machine capability, so that when the technology shifts, a regulation lands, an access regime changes, or a client situation exceeds the tools, the organisation can still respond. Resilience requires redundancy, and redundancy has a cost that efficiency accounting undervalues until the day it is the only thing that matters.

Section 10. The Stitching

Human judgment is the stitching in the firm's work: invisible while everything holds, easy to mistake for something the process no longer needs, and the first thing exposed when the seam gives. Nobody inspects a well-made garment and admires the thread. The fabric gets the attention, and now the fabric, the fluent output, the polished deliverable, has become nearly free. An organisation that responds by unpicking its stitching, the judgment of its reviewers, the apprenticeship of its juniors, the translation of its managers, will look for a while like it has lost nothing at all. Every metric will say the garment is intact. Then the seam takes a real strain: a crisis, an examination, an outage, a client question the machine cannot answer, and what shows is not a flaw in the technology but the absence of everyone who would once have caught it.

The anger in this paper, and I am aware there is some, is not directed at the people making these decisions. It is directed at a workflow design that was adopted because it photographed well in a board pack, and that quietly converts professionals into bystanders. That design is a choice, and it can be unchosen. Elsewhere in this series I have argued that AI amplifies what an organisation already has, and I have described a method for making AI-assisted judgment durable and transferable between people. This paper needed neither to stand up: its argument is complete on its own terms. The most common deployment pattern erodes exactly the thing most worth amplifying, while binding the organisation ever more tightly to systems it does not control. The stitching can be strengthened. But first you have to stop cutting it.

A Note on Sources

The claims in this paper rest on three different kinds of ground, and the reader is entitled to know which is which.

On the Cobra Effect: the Delhi bounty story has no known primary source. It circulates largely through Siebert (2001), and I present it as an illustrative parable, possibly apocryphal, rather than as documented colonial history. The mechanism it illustrates is separately established: Goodhart's original formulation appears in a 1975 paper on monetary policy (Goodhart, 1975), and the popular phrasing, that when a measure becomes a target it ceases to be a good measure, is in fact Marilyn Strathern's later paraphrase (Strathern, 1997). The line about history humming the same tune is my own phrasing of a sentiment often attributed, without verification, to Mark Twain.

On the entry-level employment evidence: the Stanford payroll study (Brynjolfsson, Chandar and Chen, 2025) is a working paper, and its findings are actively contested in the way early findings should be. The authors' own February 2026 follow-up concedes that under the most stringent controls the divergence becomes clearly significant only from 2024, and explicitly warns against treating AI as the sole cause. I use it as evidence consistent with the mechanism this paper describes, not as proof of it. The Deloitte and EY intake reductions and the entry-level vacancy figures were checked against press coverage (Myers, 2025); the press reports the firms' numbers rather than auditing them, and neither have I. Larger reduction figures circulating for other firms could not be verified against an accessible source, so I have not used them.

Claims resting on my own practice rather than published evidence: the prevalence and character of shadow AI use, including the motives I attribute to it; the account of how reviewers actually behave inside approval chains; the client-development function of the graduate intake as I have observed it, though the general alumni mechanism is documented (Dachner and Makarius, 2021); the Big Four conversation, which took place at a conference in 2025 and which I report anonymised and from memory; and the 10/80/10 banding, which is a practitioner's heuristic, not a research finding.

Arguments that are inference rather than demonstration: that the talent pipeline damage will surface at the severity and on the timescale I suggest; that the trust-mechanism account of the firm is the right reading of where professional services value is moving; that supervisory attention will extend critical-third-party treatment to AI model providers. I believe all three; I cannot yet prove any of them, and I have tried to say throughout what I would need to see to change my mind.

The regulatory claims were checked against the text of Regulation (EU) 2024/1689 as published in the Official Journal, including the phased application dates and the Annex III high-risk categories, and against Regulation (EU) 2022/2554 on digital operational resilience. The account of the Framework for Artificial Intelligence Diffusion and its rescission was checked against the Bureau of Industry and Security's publications in the Federal Register. Application dates and trade rules can be amended, and readers should confirm the current state of each before relying on it.

References

Bainbridge, Lisanne. Ironies of Automation. Automatica, volume 19, number 6, 1983.

Beane, Matt. Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail. Administrative Science Quarterly, volume 64, number 1, 2019.

Beane, Matt. The Skill Code: How to Save Human Ability in an Age of Intelligent Machines. HarperBusiness, 2024.

Brynjolfsson, Erik, Chandar, Bharat and Chen, Ruyu. Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. Stanford Digital Economy Lab working paper, November 2025.

Coase, Ronald H. The Nature of the Firm. Economica, volume 4, number 16, 1937.

Dachner, Alison M. and Makarius, Erin E. Turn Departing Employees into Loyal Alumni. Harvard Business Review, March–April 2021.

Elish, Madeleine Clare. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, volume 5, 2019.

European Parliament and Council. Regulation (EU) 2022/2554 on Digital Operational Resilience for the Financial Sector (DORA). Official Journal of the European Union, 2022.

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

Bureau of Industry and Security. Framework for Artificial Intelligence Diffusion. Federal Register, 90 FR 4544, January 2025. (2025a)

Bureau of Industry and Security. Rescission of the Framework for Artificial Intelligence Diffusion. Federal Register, May 2025. (2025b)

Goodhart, Charles A. E. Problems of Monetary Management: The UK Experience. In Papers in Monetary Economics, volume 1. Reserve Bank of Australia, 1975.

Huy, Quy Nguyen. In Praise of Middle Managers. Harvard Business Review, September 2001.

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

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

Myers, Connor. As If Graduating Weren't Daunting Enough, Now Students Like Me Face a Jobs Market Devastated by AI. The Guardian, 6 July 2025.

Lave, Jean and Wenger, Etienne. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, 1991.

Nonaka, Ikujiro and Takeuchi, Hirotaka. The Knowledge Creating Company. Oxford University Press, 1995.

Plato. Phaedrus. Translated by Alexander Nehamas and Paul Woodruff. Hackett Publishing Company, 1995.

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

Siebert, Horst. Der Kobra-Effekt. Deutsche Verlags-Anstalt, 2001.

Strathern, Marilyn. Improving Ratings: Audit in the British University System. European Review, volume 5, number 3, 1997.

Taleb, Nassim Nicholas. Skin in the Game: Hidden Asymmetries in Daily Life. Random House, 2018.

Zuboff, Shoshana. In the Age of the Smart Machine: The Future of Work and Power. Basic Books, 1988.